PERIODICAL ROOM
GENERAL LIBRARY
UNIV. OF MICH.

NOV 22 1934

The JOURNAL of
EXPERIMENTAL EDUCATION

Volume III | SEPTEMBER, 1934 | Number 1

CONTENTS

A Configurational Interpretation of Memory as Related to the Learning Process: ROGER M. BELLOWS 1

Statistical Techniques for the Study of Institutional Differences: WILHELM REITZ 11

The Variation of High School Senior and College Freshman Classes: VERGIL E. LYON 25

Five Years of Experience with a Remedial Reading Course for College Students: J. R. GERBERICH 36

Intercorrelations of Arithmetical Abilities: LEO J. BRUECKNER 42

Results of Use of Machines for Testing and for Drill, Upon Learning in Educational Psychology: JAMES KENNETH LITTLE 45

Elements of Chord-Reading in Music Notation: OTTO ORTMANN 50

$5.00 A YEAR | PUBLISHED QUARTERLY | $1.50 A COPY

EDWARDS BROTHERS, INC.
LITHOPRINTERS AND PUBLISHERS
ANN ARBOR, MICHIGAN

Application filed for entry as second-class matter. Printed in U. S. A.

EDITORIAL BOARD

JOURNAL OF EXPERIMENTAL EDUCATION

A. S. Barr, Chairman, Professor of Education, University of Wisconsin, Madison, Wisconsin.

Carter V. Good, Professor of Education, University of Cincinnati, Cincinnati, Ohio. Editorially responsible for materials on supervision and psychology of learning and teaching.

Henry Harap, Associate Professor of Education, Western Reserve University, Cleveland, Ohio. Editorially responsible for materials on experimental studies of curriculum construction.

Walter S. Monroe, Professor of Education, University of Illinois, Urbana, Illinois. Editorially responsible for materials on measurements, statistics, and methods of experimental research.

George D. Stoddard, Director, Child Welfare Research Station, State University of Iowa, Iowa City, Iowa. Editorially responsible for materials on child welfare, guidance, and development.

CONTRIBUTING EDITORS

Harry J. Baker, Director, Psychological Clinic, Detroit Public Schools, Detroit, Michigan.

W. E. Blatz, Nursery School Division, University of Toronto, Toronto 5, Canada.

William F. Book, Head, Department of Psychology and Philosophy, Indiana University, Bloomington, Indiana

Fowler D. Brooks, Head, Departments of Education and Psychology, DePauw University, Greencastle, Indiana.

William A. Brownell, Professor of Educational Psychology, Duke University, Durham, North Carolina.

Leo J. Brueckner, Professor of Education, University of Minnesota, Minneapolis, Minnesota.

Herbert B. Bruner, Professor of Education, Teachers College, Columbia University, New York City.

Barbara S. Burks, Psychologist, Institute of Child Welfare, University of California, Berkeley, California.

Otis W. Caldwell, Professor of Education, Teachers College, Columbia University, New York City.

Ellsworth Collings, Dean, College of Education, University of Oklahoma, Norman, Oklahoma.

Philip W. L. Cox, Professor of Secondary Education, New York University, New York City.

Edgar A. Doll, Director of Research, Training School, Vineland, New Jersey.

Harl R. Douglass, Professor of Education, University of Minnesota, Minneapolis, Minnesota.

Jack W. Dunlap, Associate Professor of Education, Fordham University, New York City.

Paul H. Furfey, Professor of Psychology, Catholic University of America, Washington, D. C.

Florence L. Goodenough, Professor, Institute of Child Welfare, University of Minnesota, Minneapolis, Minn.

M. E. Herriott, assistant director Division of Psychology and Educational Research, Public Schools, Los Angeles, California.

Karl J. Holzinger, Professor of Education, University of Chicago, Chicago, Illinois.

L. Thomas Hopkins, Curriculum Specialist, Lincoln School of Teachers College, 425 West 123rd Street, New York City.

C. L. Huffaker, Professor of Education, Bureau of Educational Research, University of Oregon, Eugene, Ore.

Kai Jensen, Assistant Professor of Education, University of Wisconsin, Madison, Wisconsin.

Harold E. Jones, Director of Research, Institute of Child Welfare, University of California, Berkeley, California.

Edward A. Lincoln, Assistant Professor of Education, Graduate School of Education, Harvard University, Cambridge, Massachusetts.

E. F. Lindquist, Professor of Education, State University of Iowa, Iowa City, Iowa

Lois Hayden Meek, Professor of Education, Director, Child Development Institute, Teachers College, Columbia University, New York City.

C. W. Odell, Associate Professor of Education, University of Illinois, Urbana, Illinois.

Willard C. Olson, Director of Research in Child Development, University of Michigan, Ann Arbor, Mich.

W. E. Peik, Associate Professor of Education, University of Minnesota, Minneapolis, Minnesota.

S. L. Pressey, Professor of Educational Psychology, Ohio State University, Columbus, Ohio.

W. H. Pyle, Professor of Educational Psychology, Detroit Teachers College, Detroit, Michigan.

Paul T. Rankin, Supervising Director, Board of Education, Detroit, Michigan.

Clarence E. Ragsdale, Assistant Professor of Education, University of Wisconsin, Madison, Wisconsin.

H. H. Remmers, Director, Division of Educational Reference, Professor of Education and Psychology, Purdue University, Lafayette, Indiana.

G. M. Ruch, Professor of Education, University of California, Berkeley, California.

Earl U. Rugg, Head, Department of Education, Colorado State Teachers College, Greeley, Colorado.

Peter Sandiford, Professor of Educational Psychology, Director, Educational Research, University of Toronto, Toronto, Canada.

Douglas E. Scates, Director, School Research, Cincinnati Public Schools, Cincinnati, Ohio.

Raleigh Schorling, Professor of Education, Supervisor, Directed Teaching and Instruction in University High School, University of Michigan, Ann Arbor, Michigan.

Mandel Sherman, Associate Professor of Education, University of Chicago, Chicago, Illinois.

Helen Thompson, Research Associate, Yale University, New Haven, Connecticut.

Herbert A. Toops, Professor of Psychology, Statistics and College Personnel, Ohio State University, Columbus, Ohio.

T. L. Torgerson, Assistant Professor of Education, University of Wisconsin, Madison, Wisconsin.

M. R. Trabue, 603 East Franklin Street, Chapel Hill, North Carolina.

Douglas Waples, Graduate Library School, University of Chicago, Chicago, Illinois.

Beth L. Wellman, Research Associate Professor, Child Welfare Research Station, State University of Iowa, Iowa City, Iowa.

The JOURNAL *of*
EXPERIMENTAL EDUCATION

Volume III | September, 1934 | Number 1

A CONFIGURATIONAL INTERPRETATION OF MEMORY AS RELATED TO THE LEARNING PROCESS[1]

by
Roger M. Bellows
The Ohio State University

I. HISTORICAL ASPECTS

Our dispositions are so permeated and set by educational, social, and language mores that the notions as to what constitutes memory are definitely grounded in past usage, and in obsolete custom insofar as an objective treatment of this problem is concerned. It is as a result usually difficult for the average reader to break away from these bonds of social heritage. A foreword may save labor by expressing purposes and disclaiming unwarranted inferences. A few pages further on the reader will find a more or less radical departure from the customary conception of this psychological problem. The memory enigma and kindred problems are formulized as growing out of the mere fact that systems of differentiated protoplasm are related to environments. The problem of memory, and of learning also, are fundamentally resultant properties, or supra-summative products of this relation.

The classical notions as to the nature of memory and its physiological correlate, out of which our common concepts of the process eventuated, are many and varied. Perhaps the earliest clear-cut theory of memory is the doctrine of transmigration of the Pythagoreans. Pythagoras believed that he remembered things which had occurred in a former state of existence. He considered memory a function or faculty of the soul. The soul was a storehouse for events which occurred in a former existence. Its function could continue beyond the limits of the present life. His theory is quite similar to, and anticipates, Plato's doctrine of reminiscence.

According to Burnham (4) Plato suggests as a definition of memory "the power which the soul has of recovering when by itself, some feeling which she experienced when in company with the body." In one place Plato likens the soul to a waxen tablet which may be writ upon by experience. Socrates calls this conceptural example a "waxen figment." Zeno, the Stoic, is said to have adopted this scheme. In this view is the idea that our empirical experience plays a part and that the environment is influential in our psychological processes. For Plotinus memory is a "faculty of animals, men, and angels." These early animistic writers were impressed by the fact that native or genetic forces somehow play a part in our experiences.

The ancient thinkers were not slow to guess that some physical or extended substance might be the explanation of the puzzle. Diogenes of Appolonia "explained forgetting by stating that it was due to an arrest of the equal distribution of air throughout the body, because he noted that easier breathing follows the recall of what was forgotten" writes Burnham (4). Parmenides, the Eliatic, thought the problem

1. The writer is indebted to Dr. R. D. Williams for his helpful interest and suggestive criticisms.

rested upon the way the light and heat and the dark or cold are distributed in the body. The physiological correlate of memory was seriously considered by the Greek philosophers.

During the seventeenth century Descartes clearly severed the mind from the body in his psychophysical interactionism. In Descartes is to be found the basis of the present-day theories of memory. While memory is a function of the soul for Descartes it depends upon the extended physical substrate for its initiation as a mental or soul act. His theory of memory is similar, in a somewhat gross way, to the doctrines of conditioning of our contemporary theories, which depend upon the mechanistic postulate of preferential nerve pathways. Descartes' concept of enlargement of the pores of the brain by use is comparable to our notion of the reduction of synaptic resistance due to a growth or electrochemical change in the endings of the nerve processes at the synapse. Examples of such notions are Kapper's "law" of neurobiotaxis, Bok's of stimulogenous fibrillation, and the Helmholtzian double layer ionization theory. These theories all attempt to account for memory on the basis of the development of permeability of discrete nerve pathways. They are atomistic in as much as they hold that memory is a result of the growth of parts.

For Descartes memory depended upon two things: innate disposition and impression. The innate faculty notion was largely discarded by the advent of the English associationism beginning with Hobbes and Locke. Of all the extent theories which impregnate thought upon this problem today associationism is perhaps the most virile. These philosophers took over the atomic dogma from the Newtonian physics. They assumed that the mind was made up of elements, and that to explain the adherence or association of these atoms was to explain memory. They were interested, accordingly, in the force which held these mental elements (which were at first ideas, and later sensations and images) together. This force was named. It was called the force of association. The English systems of psychology, except those of such writers as Ward and Stout, were built around this mysterious "force." Even Hume yielded to this type of interpretative principle in his concept of "gentle force."

The historical notions of memory might be divided into three parts. First we have those which insist that memory is a manifestation of some innate psychic faculty, or spiritual entity. Next we have those views which postulate an adhesiveness of elements, or the atomic associationistic schema of the English associationists. To these two views we note also an extension, in contemporary theory, of the atomic view. This standpoint is based upon the physiology of neuromuscular units.

The extension of the atomic scheme came about as a result of Spencer's emphasis upon reflex units in the nervous system, and was enhanced by the Sheringtonian and Pavlovian evidence for mechanistic conditioning. As a result of this emphasis upon reflex units, two present-day schemes of explaining retention were coined. The simplest of these theoretical trends is the Watsonian frequency-recency doctrine. The other theory of "stamping in" neural bonds is the Bentham-to-Thorndike pleasure-pain doctrine.

Watson was perhaps the most influential exponent of the formation of preferential nerve pathways by repetition. He states, "So, instead of speaking of memory, the behaviorist speaks of the retention of a given habit in terms of how much skill has been retained and how much has been lost in the period of no practice. We do not need the term 'memory', shot through as it is with all kinds of philosophical and subjective connotations" (33, p. 179). He "explains" memory on the basis of verbal and manual habits which have been formed toward particular sets or pairs of objects or persons. The development of these habits of bodily and verbal responses (postured responses) is assumed to take place by the mechanistic or atomistic theory of conditioning. Due to our original acquaintance (his practice period) with an object or person, the repetition of a verbal stimulus would "re-establish the whole of the old habit." For Watson memory is, then, the retention of verbal or manual habits,

i.e., learning by conditioning, or the formation of preferential pathways in the nervous system, assuming these pathways to be discrete and specific in nature.

The doctrine that pleasure is the prime factor of motivation of all human activity had its origin in the writings of Aristippus and Epicurus.[1] Bentham and J. S. Mill are its more recent exponents. With Spencer the same doctrine was framed in, and apparently fortified by, the thesis of differential nerve impulse conduction over reflex units. Spencer maintained that pleasantness was an accompaniment of increased nervous activity and that pain accompanies decreased nervous activity; and that success, being pleasant, caused a greater permeability to subsequent diffuse nervous discharge at the time of the later trial. Thus retentivity is said to be a result of the pleasantness of the previous event. A. Bain and also J. M. Baldwin accepted this doctrine practically as formulated by Spencer, according to Cason (6). The present-day vestiges of Thorndike's alleged law of effect are rooted in this historical tradition.

II. THE PROBLEM RESTATED

(a) The role of the organism

The views of memory have come more and more to point to the protoplasmic substrate as the fundamental basis of the phenomenon, and ever away from a spiritual or mentalistic scheme. Further, the theories of memory have taken on a more definite aspect in regard to the particular physiological parts, or mechanisms, involved.

A question here arises concerning the ultimate scientific value of this latter trend. Can one consider memory as resulting from the activity of parts as such, or is memory something over and above mere physical stuff? Many writers appear to think that this is the case. Some contemporary psychologists even go to the extent of postulating hormic and spiritual schemes comparable with those of the early Greek thinkers, because there is an indication that physiological processes per se do not account for the phenomenon. On the other hand most present-day writers hold to the belief that the process is a result of the activation of nervous pathways, and traces in the neural system. While this standpoint has gained in popularity it is perhaps no more productive than the early animistic dogma in lighting the way toward an insight into the nature of this psychological process.

Memory cannot be explained upon the basis of anything we now know about the nervous system and its function. Lashley insists that we know little more of the action of the nervous impulse than did Descartes (almost three hundred years ago and nearly two centuries before the formulation of the cell doctrine) when he described the action of his animal spirits.

The experimental evidence of the work of both Franz and Lashley indicates a lack of specificity in function of neural processes. The organismic thesis appears to be coming more and more into its own. Lashley's work indicates the vicariousness of nervous tissue (20). "It is doubtful whether the conception of limited conduction paths through the gray matter can be applied even in the case of the spinal reflexes" (19, p. 561). In the same place in regard to Descartes' notion of enlargement of the pores in the brain by the passage of his animal spirits Lashley says, "His statement sounds remarkably like the reduction of synaptic resistance by the passage of the nerve impulse." In this connection Kantor says, "As we have had occasion to see, ... the neural theory advanced to support and make plausible the revivals and storehouses has contributed nothing solid to psychological science" (15, p. 117). In opposition to the doctrine of discrete function of the nervous system Irwin insists that "the reflex chain theory of the physiologist did not seem to fit the development of specificity of activity" (13, p. 128).

1. For a general historical exposition of the hedonistic view the reader may be referred to Lecky's History of European Morals, Ch. 1, and to Watson's Hedonistic Theories from Aristippus to Spencer.

(b) The term--"memory"

The term memory is an abstraction. Memory is not a tangible entity. It results as a psychological process, from the interfunctioning of the organism with the forces of the environment, but it cannot be found within the organism or in the environment. It does not exist as a part of the organism nor does it eventuate from the mere activity of a part or parts of the protoplasmic system which is involved.

To exemplify the thought of the above paragraph one might draw an analogy from the word "health." Parts of the body alone cannot make up health because the term is also, like memory, a linguistic abstraction. The connotation of the word results from social usage. Every organ of the body might be sound and fit to do its work, but cannot do so because of faulty integration. Health is a state of togetherness of function. It eventuates when all the forces which are influential in bringing about the abstract property of living things--which we have named health--are working in harmony. Health is not to be found in the parts, but in the functioning whole together with all the natural forces which interact in bringing about this condition.

In the abstract term "memory" is the connotation of a carrying over in time of a psychobiological configuration. One cannot designate the stimulus, either external or internal, as being of value in describing this carry-over in time from one trial or event to the next. The word stimulus is ambiguous in connection with such a psychological problem as it is an artifact, another part which has been abstracted from a total psychological process. The stimulus or the stimulating situation cannot be considered as being separate from the total process. The line of demarcation between stimulus and response cannot be pointed out. Neither is one organ or system of parts (e.g., the nervous system) important by itself, nor is the organism of itself of sufficient importance in a psychological process to account for complex psychological phenomena. Rather the organism must be considered as it is related to the forces of the environment. Upon this relation the problem of memory depends.

(c) The nature of the psychological process

Memory is a supra-summative property of living and growing protoplasm. This property depends upon the organismic-environmental relation. As such it is not necessary to employ an innate faculty idea to describe and explain the process. Neither is it desirable to adhere to a system of elementarism or atomism in an adequate statement of the problem.

Memory depends upon the organismic-environmental relation. This statement involves several interesting facts. (1) On the side of the organism memory depends upon a physiological state. The physiological correlate of memory, which is the fact of retentivity, might be defined as a particular degree of physiological differentiation of the organism as a whole. (2) Memory itself is defined as a residual dynamic relation, which is a higher level of physiological growth, which exists after the partial or complete resolution of an organismic-environmental tension. (3) Retentivity or memory is one condition for the learning process. Let us see what is included in these three statements.

Memory may be said to depend upon a physiological state of degree of physiological differentiation of the organism as a whole. Without a sufficiently mature animal there could be no memory process for a given situation, for tissue would not be so differentiated and irritable, and integration would not be sufficient to bring about the organismic state of dynamic relation. In the older nomenclature the organism would not have an experience or perception. Amoebae cannot remember a sound, for example, because tissue has not evolved to a sufficiently high level of specialization so that the sound affects this organism; that is, it does not bring about a condition of dynamic relation. In the human being tissue is so evolved and irritable, and integration is relatively so complete that many sounds can set up this condition of stress-strain.

The more irritable the tissue of the organism the more diverse are its experiences. More of the energy changes of the environment influence the protoplasmic

system. A three year old child is said to "forget" his house number. The vestigial dynamic relation is not present in the latter case, either because no such relation was formed due to the fact that the degree of physiological differentiation and integration was not sufficient to bring about this state of organismic imbalance, i.e., "house number" was not experienced, or because the residual dynamic relation was so weak as not to enable the child to recall the house number.

Ability to recall in the experimental usage is the criterion of the psychological act of remembering, so that we say no remembering has taken place if the subject cannot recall the material. In regard to recall as a criterion of memory, and in support of the notion of vestigial dynamic relations as a description of memory, cases of unusual recall under extraordinary organismic-environmental relations are in point. Warren (32) reported the case of an elderly man who returned to his Alma Mater to attend the commencement exercises. Under the strain and excitement of the experience he arose upon occasion and recalled word for word his freshman oration which he had delivered many years before, and which he had not recalled since. In the writings of Freud and his followers are other startling examples of latent memory.

The conditions of memory may be said to be the same as those which enable the individual to be placed under psychobiological tension or undergo a condition of dynamic relation. Without a state of dynamic relation no memory would be evidenced in behavior by recall or ability to recognize, the latter being a common criterion of memory. To explain the "why" of memory is to state the conditions of the formation of this organismic-environmental tension. We shall now examine the basic conditions of memory, or the stage when it is set for this psychological process.

In the first place the subject, either animal or man, must have attained a certain degree of physiological maturation. This degree of irritability of tissue is the fundamental condition of memory. When we recall that memory is an abstraction it is evident that in one sense this basic condition for the process is the physiological correlate of the process itself. That is to say, by movement and the life processes animals grow and become differentiated in time. By this very growth animals are enabled to be influenced by more and more of the energy forces of the environment which surround their sensitized tissue.

By virtue of differentiation and specialization of tissue organisms are potentially under a strain, or have dynamic relations, at all times. That is, they have potential unsatisfied needs.

This statement simply implies that animals have more experiences when they are more differentiated. Animals low in the phylogenetic scale, such as protozoa, are influenced by environmental forces. Amoeba Proteus shows, upon observation, phototropic and thigmotactic behavior due to changes of light intensity and pressure of the environmental envelope of their surface membranes. Iron, too, rusts under certain environmental conditions. But these systems are relatively simple and of low grade of differentiation when compared with the human. They cannot be influenced by changes of 50 double vibrations in the sounds made by vibrating tuning forks, for example. In the case of the human and many of the infra-human organisms much smaller changes in the energy of the environmental forces which play upon their specialized tissue will bring about activity which indicates the presence of a strained condition of dynamic relation. Dynamic relations cannot be set up without differentiation of tissue, and the degree of this physiological growth must be adjusted to the relative intensity and kind of energy change.

There are three problems which are related closely to that of memory, and which may be interpreted from the viewpoint of configuration, or dynamic relations. One of these is vicarious selection--the fact that one event will be remembered instead of another, or that one movement will be retained in patterned behavior while another will drop out. A second problem, which in part accounts for vicarious selection, is that of motivation. The third deals with the relation of memory, as here conceived, to the learning process.

The problem of vicarious selection. Shortly after the beginning of the nineteenth century Thomas Brown was concerned with the fact that some things would be remembered when others would be forgotten. He sought the reasons for this fact. Finally he postulated additional or secondary "laws" of association, to account for it. He said that those experiences which were originally most intense, frequent, recent, and vivid would take precedence in the stream of mental states at the time of recall over those which were less so. He still dealt with the intangible, abstract "force" of association. In regard to vicarious selection as improvement of behavior, Cason and others have criticized frequency and recency as a basic explanatory factor (5, 16, 17, 18, 34). Cason has recently contributed a rather complete summary of the defects of the "law of effect" based upon experimental evidence (6).

Complex bio-social factors influence dynamic relations in the case of the more differentiated human subject. The fact that a subject will remember one person or event and not another is dependent upon the first or fundamental condition of memory treated above, and also upon motivational influences. These latter factors might be referred to as determiners of degree of sureness of recognition and recall.

Motivation. In the factor of motivation is included the fact that all organisms have potential needs. Adams, for example, states that needs "are universal properties of protoplasm" (1). These needs may be divided into two classes. They may be primary biological needs, e.g., food and oxygen; or they may be secondary or social. The primary needs are factors influencing the behavior of all organisms. Due to the fact that the higher animals live together and communicate with one another, the more intricate societal type of need is in evidence. Cooperative life results from the biological fact of growth and differentiation--tissue in the instance of the higher organisms has evolved to such a specialized sense-organ status that intercommunication is possible.

Examples of the social type of need are the following: approbation, information, health, wealth, property, power, reputation, the aesthetic, etc. Biological needs may conflict with social ones. Sex motivation is often contrary to social mores and customs. Due to this conflict of needs of different types is brought about a condition of organismic-environmental relation in which the primary needs are cancelled by social forces, or what is commonly called social pressure (cf. Taylor (29) on "alternative response"; and Freud on sublimation). The different types of needs interact with one another. The net result of all of the needs in interaction is the influencing motivational factor. Motivation is a contributing factor in the formation of dynamic relations.

Such a scheme of opposed or contradictory needs accounts for the fact that in cases of pathological subjects forgetting of obvious names and events takes place. The Freudians have assumed that such lapses of memory in these cases are caused by "repressions" due to social pressure. They, however, hold the unnatural "cellar" doctrine of the unconscious, with their myth of censorship of ideas which attempt to enter the conscious realm. Herbart and also Schopenhauer held similar notions regarding the suppression of ideas. This mysterious interpretation of the mind is unnecessary and lacks the virtue of parsimony when the "mind" might be considered as a potential organismic property of protoplasm which comes into being by reason of the fact that the organism is related to the environment in a peculiar manner. The major psychological processes are developmental problems. These processes result from the formation of dynamic relations.

The relation of memory to learning. When such organismic tensions are resolved, learning takes place. Learning is a reorganization of potential, or reorganization of the psychobiological field, which reduces the dynamic relation to a minimum. To make this somewhat vague paragraph clear, the fundamental differences between memory and learning will be pointed out.

The learning process is dependent upon degree of physiological differentiation, and environmental barriers to the satisfaction of needs. Differentiation is an

important prerequisite for learning for two reasons: (1) If the degree of differentiation is not sufficient, the organism cannot be subject to a particular organismic situation of tension (i.e., no problem will be present, or no dynamic relation will be effective). (2) If the organismic-environmental stress is present, differentiation is necessary to resolve the strain or reorganize the field. In case the relation is present but not partially resolved due to a lack of a sufficient level of physiological differentiation, either failure of the adaptive response, maladjustment to the situation, or death (in the instance of "jungle environment") ensues. Dynamic relations or psychobiological tensions are susceptible to experimental quantification and measurement. In the comparative laboratory, for example, Warden has measured tensions by the obstruction method (31) in the case of the sex and hunger motivated albino rat.

Obstructed needs set the stage for the learning process as they do also for the process of memory, provided that the physiological level is of a sufficient grade to meet the two conditions stated in the paragraph just above. The memory process, of which the physiological correlate is differentiation of the organism as a whole, accounts for asymptotic behavior or perfection of the act, commonly referred to as selection of the adaptive patterns of activity and elimination of the incorrect responses.

At each successive dynamic relation the maturational level of the organism is on a higher plane. That is to say, organisms may grow as they are subjected to dynamic relations. Psychological growth is a matter of residual psychobiological tensions in the organismic sense, and in this sense memory is synonymous with growth. Adams conceived of retentivity as "a universal property of protoplasm" (1), and psychological growth as so conceived is likewise a universal biological fact. Differentiation of the organism as a whole thus enables animals to acquire more contacts with nature, and thereby subjects them to more diverse dynamic psychobiological relations. When the white rat meets a new maze situation it is subjected to a new organismic tension--a condition of distortion of the psychobiological field. The first acquisition of the true response is a reorganization of this field of forces, or a partial resolution of the condition of dynamic relation of genetic-environmental forces. This acquisition may be conceived as leaving a residuum of physiological forces within the integral organism which coalesce with one another. Learning may be considered as evolution of psychobiological tensions, as has been suggested by a number of Gestalt psychologists, or the process may be defined as perceptual evolution.[1] Memory, or the propensity for recall, is a process which has been abstracted out from the more basic psychological learning process. The relation of memory and learning is as a by-product of an important manufactured commodity to the basic product itself.

(d) The configurational viewpoint experimentally considered

In human subjects there is a high correlation between speed of learning and retention. This fact has been indicated experimentally by Norsworthy (22), Woodworth (36), and Pyle (26). In psychological processes Hollingworth states "correlation rather than compensation is the developmental rule" (12, p. 27). This fact is evident when one conceives of the organism as growing and differentiating organismically as a unit, rather than postulating a separate mental faculty for the various abilities, or considering the organism mechanically as built up out of a number of discrete physiological units or atoms.

It has been found in the experimental laboratory that, if after learning and before recall, certain types of activity are interpolated, this activity affects, and in some cases inhibits, the performance of recall. Müller and Pilzecker pointed out the fact of retroactive inhibition at the beginning of the century (21). They attributed the effect of interpolated activity to its

1. Cf. A former study by the writer dealing with a configurational conception of the learning process (3).

interference with the process of "neural set." The latter process was presumed to account for retention. The neural doctrine has since been criticised and largely abandoned as a result of Lashley's work in experimental physiology referred to above. But the fact of retroactive inhibition remains. It appears to fit in well with the viewpoint now being considered.

If the organism be conceived as having attained a particular degree of differentiation which enables it to undergo dynamic relations of a particular level, the resolutions of these tensions must result in the augmented physiological differentiation in a particular direction. Activities which result from different dynamic relations such as interpolated organismic-environmental situations would naturally cause differentiation in an opposite direction. The interpolated situation would thereby tend to inhibit resolution of tensions which would normally result from relations similar in nature to the former ones.

III. SUMMARY

"Memory"--the term is an abstract artifact which has no definite connotation. Its historical interpretations fall under three categories:

1. Argumentative, which assumes memory phenomena to be manifestations of some mysterious psychic faculty,

2. Atomistic associationism, which postulates an equally mysterious "force" said to account for the memory process, and

3. Physiological atomism, which insists that the psychological evidences of memory are due to traces or grooves left by experience in discrete neural parts of the organism.

None of these schemes is adequate in an objective custom-free working conception of this psychological problem.

Rather the phenomenon results from the fact that organisms are related in a particular manner to environments. This fact entails the following considerations:

1. Memory depends upon one fundamental condition. This condition is a particular degree of physiological differentiation of the organism as a whole. This is a necessary prerequisite of the process for two reasons,

a. So that the irritable organism may be subject to a state of dynamic relation. Dynamic relations are defined as states of organismic-environmental tension.

b. So that this dynamic relation may be resolved.

2. Memory is related to the learning process as the resolution of the state of dynamic relation is related to the augmented level of physiological process which eventuates from the formation and resolution of dynamic relations. Its physiological correlate is increased bodily differentiation of the organism as a whole.

3. Memory accounts for continued improvement in behavior (asymptotic performance).

4. Vicariousness of recall (the fact that one name or event is recalled instead of another) is due to the motivational factors present at the time of the formation of the particular dynamic relation, interpolated organismic-environmental conditions, and the dynamic relation existing at the time of recall. Motivation is considered as a state of obstructed need. Needs are potential universal properties of protoplasm. They may be either primary biological or societal in character, or a fusion of the two types.

5. The present fund of experimental factual evidence concerning correlation between learning and retention, and retroactive inhibition, is not contrary to the configurational scheme of the memory process.

BIBLIOGRAPHY

1. Adams, D. Q. "A Restatement of the Problem of Learning", British Journal of Psychology, xxii (1931), pp. 150-178.

2. Angell, J. R. "Review of Recent Literature on Memory", Psychological Review, 1 (1894), pp. 435-438.

3. Bellows, R. M. "Learning as Perceptual Evolution", Psychological Review, xl (1933), pp. 138-159.

4. Burnham, W. H. "Memory Historically

and Experimentally Considered", American Journal of Psychology, ii (1888), pp. 225-270.

5. Cason, H. "Criticism of the Laws of Exercise and Effect", Psychological Review, xxxi (1924), pp. 397-417.
6. Cason, H. "The Pleasure Pain Theory of Learning", Psychological Review, xxxix (1932), pp. 440-466.
7. Crile, G. W. "An Electro-Chemical Interpretation of Memory", Proceedings American Philosophical Society, lxix (1930), pp. 359-368.
8. Curti, M. W. Child Psychology (New York: Longmans, Green and Company, 1930), pp. viii + 527.
9. Dunlap, K. "The Biological Basis of the Association of Ideas and the Development of Perception", Psychobiology, ii (1920), pp. 29-54.
10. Edgell, B. Theories of Memory (London: University of Oxford Press, 1924), p. 174.
11. Herrick, C. J. "The Propagation of Memories", Psychological Review, iv (1897), pp. 294-297.
12. Hollingworth, H. L. Mental Growth and Decline (New York: D. Appleton-Century Company, 1927), pp. xii + 396.
13. Irwin, O. C. "The Organismic Hypothesis", Psychological Review, xxxix (1932), pp. 128-146.
14. Jennings, H. S. The Behavior of Lower Organisms (New York: Columbia University Press, 1923), pp. xiv + 366.
15. Kantor, J. R. Principles of Psychology, Vol. 2 (New York: Alfred A. Knopf, Inc., 1926), pp. xii + 524.
16. Koffka, K. The Growth of the Mind (New York: Harcourt, Brace and Company, 1929), pp. xii + 524.
17. Köhler, W. Gestalt Psychology (New York: Horace Liveright and Company, 1929), pp. xvi + 383.
18. Kuo, Z. Y. "The Fundamental Error of the Concept of Purpose and the Trial and Error Fallacy", Psychological Review, xxxv (1928), pp. 414-433.
19. Lashley, K. S. "Learning: I, Nervous Mechanisms in Learning", Ch. 14 in The Foundations of Experimental Psychology, pp. 524-563, edited by Carl Murchinson (Worchester: Clark University Press, 1929).
20. Lashley, K. S. "Studies of Cerebral Function in Learning: VI, The Theory That Synaptic Resistance Is Reduced by the Passage of the Nerve Impulse", Psychological Review, xxxi (1924), pp. 369-376.
21. Müller, G. E., and Pilzecker, A. "Experimentelle Beiträge zur Lehre vom Gedächtnis", Zeit. f. Psychol., Erzgsbd. I, pp. xiv + 300.
22. Norsworthy, N. "Acquisition as Related to Retention", Journal of Educational Psychology, iii (1912), pp. 214-218.
23. Ogden, R. M. "Learning as Improvement", American Journal of Psychology, xxxix (1927), pp. 235-258.
24. Pavlov, I. P. "The Reply of a Physiologist to Psychologists", Psychological Review, xxxix (1932), pp. 91-127.
25. Pillsbury, W. B. "The Effects of Training on Memory", Educational Review, xxxvi (1908), pp. 15-27.
26. Pyle, W. H. "Standards of Mental Efficiency", Journal of Educational Psychology, iv (1913), pp. 61-70.
27. Renshaw, S. "The Errors of Cutaneous Localization and the Effect of Practice on Localizing Movement in Children and Adults", Journal of Genetic Psychology, xxxviii (1930), pp. 223-238.
28. Scott, R. R. "Some Suggestions on Learning with the Point of View of Gestalt Psychology", Journal of Educational Psychology, xxi (1930), pp. 361-366.
29. Taylor, W. S. "Alternative Response as a Form of Sublimation", Psychological Review, xxxix (1932), pp. 165-173.
30. Titchener, E. B. "Memory and Imagination: A Restatement", Psychological Review, xix (1912), pp. 158-163.
31. Warden, L. H., and Nissen, H. W. "An Experimental Analysis of the Obstruction Method of Measuring Animal Drives", Journal of Comparative Psychology, viii (1928), pp. 325-342.
32. Warren, H. C. "Two Cases of Long Latent Memory", Psychological Bulletin, (1918), pp. 207-209.
33. Watson, J. B. Behaviorism (New York: W. W. Norton and Company, 1924), p. 251.

34. Wheeler, R. H. The Science of Psychology (New York: Thomas Y. Crowell Company, 1929), pp. xvii + 556.

35. Woodworth, R. S. "A Contribution to the Question of 'Quick Learning, Quick Forgetting'", Psychological Bulletin, xi (1914), pp. 58-59.

STATISTICAL TECHNIQUES FOR THE STUDY OF INSTITUTIONAL DIFFERENCES

by
Wilhelm Reitz
University of Chicago

The results of testing programs, both state and city, have revealed considerable differences in scholastic aptitude, educational achievement, and other qualities of student bodies of various institutions. Often these differences are much greater than those found in the student body of a single institution. Henmon, in discussing the distributions of percentile ranks on the Ohio State psychological test obtained for colleges in Wisconsin, states:[1]

> "We are nowadays accustomed to a wide range of individual differences in any test that may be given, but are scarcely prepared for such extraordinary range of group or institutional differences as appear here. A range in medians for the 18 or 20 colleges on a basis of declaration of intention to enter from 39 - 76 and in the actual entrance from 40 - 76, goes far beyond what anyone would have expected. It is a striking fact that the upper quartile point for actual entrants in six of the institutions does not exceed the median in the two institutions with the highest selection. Again, the median percentile rank of actual entrants for the three institutions is below that of the lower quartile in two or three institutions."

Bixler, in studying the variations in average scores on the American Council Psychological Examinations among the high schools in the Chicago area, reports:[2]

> "The range of the individual school was from 4 - 334, with a possible score of 370, and the range of the median score of the class was from 65.0 to 201.4. The lowest median score was lower than the lowest individual score in six schools. The median scores in three schools were as high as, or higher than, the median scores of the 1930 freshman class of the University of Chicago."

Similar variations, with regard to intelligence, English, and achievement in other subjects, have been reported in the Pennsylvania study.[3] Haggerty obtained like results in connection with his studies on the revision of standards of the North Central Association. Placing 60 institutions in the order of their average intelligence scores calculated from data covering a three-year period, he found that the institution ranking highest had an average median score of 228 and the lowest a score of 95.[4]

There is no doubt that these differences raise some very fundamental questions with regard to the educational policies of institutions either receiving students from or supplying students to other institutions. Some of the educational problems met by institutions receiving students are briefly outlined below.

Should an institution find that a certain measure, say scholastic aptitude expressed in percentile ranks on the national basis as provided for the American Council Psychological Examination, shows a variation of the means from year to year to be less than the variation about the means[5] of the successive years, that institution would be justified in assuming that the average of the various yearly averages is characteristic

1. V. A. C. Henmon and F. O. Holt, A Report on the Administration of Scholastic Aptitude Tests to 34,000 High School Seniors in Wisconsin, June, 1931.
2. R. W. Bixler, "Rating the College Entrant", Journal of Higher Education, III (October, 1932), 361-65.
3. Study of the Relations of Secondary and Higher Education in Pennsylvania, The Carnegie Foundation for the Advancement of Teaching in Cooperation with the Joint Commission of the Association of Pennsylvania College Presidents and the State Department of Public Instruction, New York, 1931.
4. M. E. Haggerty, "The Product of Higher Educational Institutions", North Central Association Quarterly, Preprint, Table IV.
5. The expression "variation of the means" refers to an estimate of the variance (i.e., the square of the standard deviation) derived from the yearly averages; "variation about the means" refers to an estimate of the variance derived from the yearly standard deviation.

of its student body. The institution could further assume that its internal educational work in the course of time has been adapted to students of that level of ability and that the curriculum, aims and objectives, methods of instruction, evaluation of students' achievement--in brief, all institutional functions--are fairly adjusted to the particular level of ability. On the other hand, should the variation of the means be larger than the variation about the means, indicating a heterogeneous student body from year to year, the institution could not make such assumptions, but must assume that the fluctuation in student material is met by proper flexibility in its educational procedures. Should the heterogeneity exceed certain bounds, it probably would be impossible for the institution to function effectively. This might be true, for example, if the student body for one year were highly intelligent, having an average equal to the 75 percentile, and the next year would drop to the 25 percentile; or would for one year be coeducational with a fifty-fifty ratio, and the next year would consist only of male or only of female students. Unless the machinery were highly flexible and the staff able to adapt the internal work to the prevailing status of its student body, serious institutional deficiencies would result.

A further advantage accruing from the knowledge of the relative variability of the means from year to year, and about the mean of any one year with regard to certain measures, is seen in the possibility that institutions which find the variability of the means to be less than that about the mean for any one year might be able to determine rather stable quantitative norms for the selection of their students.

Thus, for example, an institution which finds that it has a student body varying during a three- or four-year period around the 75 percentile rank on the American Council Psychological Examination seems to be justified in using such a percentile, with a corresponding range, as a definite criterion for the admission of students. Several similarly determined criteria should be of considerable help to institutions in selecting and guiding their student bodies.

The value of such criteria assumes even greater importance in the light of the recent move of the North Central Association to revise its standards in judging institutional excellency. The shift from quantitative and mechanical to qualitative and judicious standards will necessitate that an institution's educational achievement be evaluated in terms of the abilities of its student body, to quote a pertinent passage by Haggerty:[1]

> "What we may legitimately ask of any institution is that it clearly apprehend its function in terms of the character of its student body, that it frankly state its purpose and its program, and that it adapt its procedures to the character of its clientele. Let any institution do these things, let it drop its pretenses of being something that in the nature of the case it cannot be, let it demonstrate that it can perform a useful service for the type of youth who seek its ministrations, and it will be as worthy of social recognition as the one which prides itself on the high intellectual level of its student group. .....
>
> "It is not unvarying conformity to a single pattern that we seek; it is effective institutional service to the widely varying abilities and interests of the young people who seek, through higher education, a meaningful preparation for the kind of life they desire and can achieve. What we require is that an institution shall illuminate its practices by a clear understanding and description of the particular functions it essays to perform for the particular group of students who choose to enter its doors. In the light of this principle an understanding of the intellectual characteristics and interests of each student group becomes of first importance to each institution itself and to any accrediting agency to which it is in any way related."

Business and industry have for many years resorted to rigid statistical techniques in the determination of the quality

1. Haggerty, op. cit., p. 9.

of raw materials and finished products;[1] but educational institutions, handling probably the biggest "business" in the country, have until recently made little progress in determining their procedures in the light of the character of their student bodies, and conversely, the qualities necessary for the achievement of their particular aims and objectives.

Although institutions have had ample data available for a number of years, adequate statistical techniques have not been sufficiently focused on them to render such data useful to studies of institutional differences as related to qualities of their student bodies. It is, therefore, the purpose of this paper to apply certain well established statistical techniques for the analysis of suitable data pertaining to the student body of educational institutions.

In essence, the techniques may be explained as furnishing an index of homogeneity. This index would determine whether or not students comprising the enrollment of an institution are so selected from year to year as to form a homogeneous or a heterogeneous population. It would also determine whether or not certain groups of students admitted year after year from a particular institution are of the "same uniform run", having due regard, of course, to individual variation, and whether or not students vary more in a particular measure than might be expected through chance.

Pertinent institutional measures will be presented first. On the basis of these data, the statistical problem will be formulated and the techniques explained. While there is no restriction to any particular measure, the percentile ranks on a national or state-wide basis, as, for example, furnished in the Manual of the American Council Psychological Examination, are used for illustration.

The whole procedure is first applied to an institution which might attempt to ascertain the central tendency of a measure pertaining to its student body, (e.g., the mean of intelligence), by defining certain of its characteristic features in the form of a quantitative expression which may ultimately serve as a standard for selective admission. A second example will show how an institution might use this technique to determine a rating of those institutions from which its students are drawn.

The following measures are presented to illustrate the first case, that of an institution attempting to determine: (1) whether or not its student material, over a period of years, has been homogeneous or heterogeneous with regard to intelligence as measured by the American Council Psychological Examination; and (2) what has been the level of intelligence of these students, with a view to determining approximately a standard and a critical range for future selective admission.

The data in Table I were prepared in the following manner. The raw scores on the American Council Psychological Examination, as reported by an institution for four successive years, were translated into the national percentile ranks as found in the Manual of Instruction. This process is necessary in order to unify[2] all distributions as though they had all been obtained on one measuring instrument. The same unification would have been obtained had the various distributions been translated by way of the national percentile ranks into a score distribution of any one year; for example, into the first year distribution. An illustration will make this clear. If the first, second, third, and fourth years represent the score distributions of the 1928, 1929, 1930, and 1931 editions respectively, and if the distributions of scores on the 1929, 1930, and 1931 editions are translated into the score distributions of the 1928 edition, the process of transforming is as follows: a score of 155 on the 1931 edition represents a national percentile rank of 55, and to a score of 146 on the 1928 edition corresponds

1. E. S. Pearson, "The Analysis of Variance in Cases of Non-Normal Variation", Biometrics, XXIII (1931), pp. 114-33.
2. These percentiles are sometimes subject to fluctuations due to the kinds of colleges reporting. For instance, the latest reports on the Psychological Examination in the Educational Record show that colleges with low median scores tend to discontinue to send in their reports on which the national percentile norms are based. This, of course, renders national norms non-comparable.

a percentile rank of 55; therefore a score of 155 on the 1931 edition is equal to a score of 146 on the 1928 edition.

This method could be used in transforming the yearly national or state distribution into one standard distribution. This is the proper method if scores are desired rather than percentile ranks.

TABLE I

DISTRIBUTIONS OF INTELLIGENCE PERCENTILE RANKS OF STUDENTS OF INSTITUTION A OVER A PERIOD OF FOUR YEARS

National Percentile Ranks	Distributions of Percentile Ranks				Total Distribution All Years
	1st Year	2nd Year	3rd Year	4th Year	
90-100	22	26	25	35	108
80-89	17	15	21	14	67
70-79	11	11	9	7	38
60-69	14	12	19	6	51
50-59	6	7	12	7	32
40-49	10	14	18	5	47
30-39	5	7	8	7	27
20-29	11	6	2	21	40
10-19	6	5	7	6	24
1-9	3	1	1	1	6
(n_i)	105	104	122	109	(N) 440
(m_i)	62.9048	65.7692	65.4098	63.3486	(M) 64.3864
(s_i)	27.5576	25.7089	24.2392	30.0920	(SD) 26.9484

The treatment of the other distributions would be similar. From a purely statistical point of view, the equating of scores of different years into the score distribution of one year used as the base is more satisfactory than the transformation of scores into percentile ranks, because score distributions more or less approach normality in distribution, while percentile ranks form rectangular distributions. Since the techniques proposed in this study are based on normal distributions, the use of percentile ranks must be thought of as a convenient measure to illustrate the whole procedure, and as permissible, insofar as errors introduced will not invalidate the results. Another way of eliminating the variations in the distributions due to the individual test editions would have been by Hull's[1] methods of rendering comparable distributions of different means and standard deviations.

After the unification of the yearly measures was obtained, the means and the standard deviations were computed for each year, as well as for the total of all years. Table I shows the following means for the various years: 62.9, 65.8, 65.4, and 63.3. For the total of all years, the means is 64.4. The standard deviations are 27.6, 25.7, 24.2, 30.1, and for the total, 26.9.

To answer the question as to whether or not the student body has been homogeneous or heterogeneous in intelligence over the four-year period, it is necessary first to determine whether or not the various means fluctuate more or less about the total mean than might be expected through chance, giving due regard to the observed variation among the individual distributions.

If the variation of the individual

1. C. L. Hull, Aptitude Testing, p. 396 ff. New York: World Book Company, 1928.

means is within chance fluctuation, the yearly distributions of students can be thought of as homogeneous or as samples of the same parent population whose statistical characteristics are approximately represented by those of the total distribution. The mean of the total distribution, therefore, would be representative of the four distributions.

Two methods will be presented and discussed by means of which the foregoing problem can be solved. One method consists in computing the correlation ratio[1] and in testing its significance. This method is fully presented by Pearson[2] in his Tables for Statisticians and Biometricians. The second method consists in a study of variance by the method Fisher[3] has outlined in his volume, Statistical Methods for Research Workers. According to the first method, a test of significance will show if the variation between the arrays (in the present study, between the various years) is significantly greater than is to be expected, without correlation, from the variation within arrays. The second method demonstrates whether the variance of the means is not more than might be expected through chance or not more than the variance about the means.

Should the test of significance be positive by either method, the yearly samples of student bodies may be considered as coming from different parent populations, and, therefore, as heterogeneous. If the test shows negative results, the samples may be considered as homogeneous; i.e., as coming from the same parent population. In the latter case, the total distribution would give a fair picture of the present population from which the institution draws its student body. Its average and the standard deviation could form the base for the determination of a standard of selective admission. Whenever the student material is heterogeneous, such a standard could not be derived in the same way, at least not with the same reliability.

By the first method, the correlation ratio (η) is computed and tested according to the procedure outlined by Pearson. This test is analogous to testing the significance of the difference between two means, i.e., by relating the difference to the standard error of the difference. In this case, if (η^2) is found, the formula is:

$$(1)\quad \lambda = \frac{\text{observed } \eta^2 - \text{mean } \eta^2}{\text{standard error } \eta^2}$$

observed η^2 = the square of the correlation ratio

$$(2)\quad \text{mean } \eta^2 = \frac{K - 1}{N - 1}$$

K = number of samples or years
N = total number of cases (students), all years
standard error η^2 is found by formula:

$$(3)\quad \frac{1}{N - 1}\sqrt{\frac{2(K - 1)(N - K)}{N + 1}}$$

The correlation ratio (η) can be found either by relating the average standard deviation of the various arrays (in this case, years) or the standard deviation of the various means of the arrays to the standard deviation of the total distribution. According to the first computation:

$$(4)\quad \eta^2 = 1 - \frac{S_1^k(n_i\, s_i^2)}{NSD^2}$$

and according to the second:

$$(5)\quad \eta^2 = \frac{S_1^k n_i(m_i - M)^2}{NSD^2}$$

In these formulas, S_1^k stands for the summation of all quantities from 1 to k, i.e., for the individual years; n_i stands for the number of students in the various years; s_i for the standard deviations of the various years; and SD and M for the standard deviation and mean of the total distribution.

1. η_{xy}: y representing the group number or the number of years, and x, the particular measure which is studied, such as intelligence.
2. K. Pearson, Tables for Statisticians and Biometricians, Part II, pp. xl and 16.
3. R. A. Fisher, Statistical Methods for Research Workers, Chapter VII, pp. 194 ff.

Institution A

In computing η^2 for Table I, it is found that

$$\text{by (4)}\quad \eta^2 = 1 - \frac{318859.77}{319534.32} = .002111$$

$$\text{by (5)}\quad \eta^2 = \frac{674.55}{319534.32} = .002111$$

The computation of η^2 by two different methods is advisable for the purpose of accuracy; the results by both methods should be identical to a reasonable number of decimal places.

The next step in the procedure consists in testing of η^2 for its statistical significance. The mean η^2 is found by formula (2). Applying this to the situation in Table I, Institution A,

$$k = 4, \quad N = 440, \quad \text{so that}$$

$$\text{mean } \eta^2 = \frac{4 - 1}{440 - 1} = \frac{3}{339} = .006834$$

The standard error η^2 is found by formula (3). Substituting the corresponding values in this formula, the following is obtained:

$$\text{Stand. error } \eta^2 = \frac{1}{440 - 1}\sqrt{\frac{2(4-1)(440-4)}{440+1}}$$

$$= .005548$$

Fortunately the mean and the standard error of η^2 need not be computed, but may be found in Table IV of Part II of Pearson's Tables for Statisticians and Biometricians. This table furnishes directly all means and sigmas of η^2 for population of 51 to 1000 and for 3 to 20 arrays. Substituting the three values thus far found in (1),

$$\lambda = \frac{.002111 - .006834}{.005548} = -\frac{.004723}{.005548} = -.85$$

it is immediately seen that the resulting ratio (λ) is negative and insignificant. The conclusion can be drawn that the yearly averages fluctuate within chance expectation having due regard to the fluctuation of the individual measures about the individual averages. In other words, Institution A has had fairly homogeneous student bodies over the four-year period studied. The total average of 64.3864 and the standard deviation of 26.9484 may be considered representative parameters of the population from which the yearly distributions have been drawn.

According to Fisher's procedure of "analysis of variance", the same conclusion can be reached, although the statistical technique is quite different. According to Fisher, the following set-up is necessary for analysis:

(6)

Variance	Degrees of Freedom	Sum of Squares
From Variation of Means	$k - 1$	$\frac{1}{N} S_1^k \{n_1(m_1 - M)^2\}$
From Variation about Means	$N - k$	$\frac{1}{N} S_1^k (n_1 s_1^2)$
Total	$N - 1$	$\frac{1}{N} S_1^N (s - M)^2$

If the hypothesis that the means differ only through chance fluctuations be true,

$$\frac{1}{N}\,\frac{S_1^k \{n_1(m_1 - M)^2\}}{k - 1} \quad \text{and} \quad \frac{1}{N}\,\frac{S_1^k(n_1 s_1^2)}{N - k}$$

will be two independent estimates of the unknown group variance σ^2. The hypothesis is tested by examining the significance of z, the natural logarithm of the difference of these two estimates of variance. To illustrate Fisher's procedure, the data pertaining to Institution A are used.

Variance	Degrees of Freedom	Sum of Squares	Mean Square	SD	$\log_e$ SD
From Variation of Means	3	1.5330	.5110	.71484	9.6643 - 10
From Variation about Means	436	724.6813	1.6621	1.28922	.2540
Total	439	726.2143	1.6542	1.28617	9.4103 - 10(z) or -.5897

Fisher's z is found to be the difference between the natural logarithms of the two standard deviations. For Institution A, this difference is 9.4103 - 10, or -.5897. The computation of z can be checked by the relation which exists between η^2 and z, namely,

$$(7)\quad z = \frac{1}{2}\log_e\left\{\frac{(N-k)\eta^2}{(k-1)(1-\eta)^2}\right\}$$

When the corresponding values are substituted in (7), the following result is obtained:

$$z = \frac{1}{2}\log_e\left\{\frac{(440-4)(.002111)}{(4-1)(1-.002111)}\right\}$$

$$= 9.4103 - 10 \quad \text{or} \quad -.5897$$

This result is identical with that obtained by the Fisher technique. Again z may be tested for its significance by relating it to its standard error,[1] which is given by the formula:

$$(8)\quad SE_z = \sqrt{\frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$n_1 = k - 1 \quad \text{and} \quad n_2 = N - k$$

Substituting the corresponding values in (8), the SE_z is found to be

$$SE = \sqrt{\frac{1}{2}\left(\frac{1}{3} + \frac{1}{436}\right)} = .4097$$

Relating z to SE_z, a ratio of $-\frac{.5897}{.4097} = -1.44$ is obtained, which signifies clearly that z is insignificant in the sense that the variance about the means is much greater than that of the means. With regard to Institution A, it may again be concluded that the variation about any one mean over a period of four years is much larger than the variation of the means from year to year, i.e., Institution A has been populated over a period of four years by students of approximately the same average intelligence.

In the correlation ratio procedure, whenever the mean η^2 is larger than η^2, there is no need to consider a test of significance of η^2, as the resulting ratio (k) would always be negative. The same is true if, in Fisher's procedure, the SD resulting from the variation of the means is smaller than that resulting from the variation about the means.

Before further discussing the foregoing results, the same technique will be applied to another institution where conditions are such that quite opposite conclusions must be drawn. The pertinent data are furnished in Table II.

TABLE II

DISTRIBUTIONS OF INTELLIGENCE PERCENTILE RANKS OF STUDENTS OF INSTITUTION B OVER PERIOD OF FOUR YEARS

National Percentile Ranks	Distributions of Percentile Ranks				Total Distribution All Years
	1st Year	2nd Year	3rd Year	4th Year	
90 - 100	7	5	22	27	61
80 - 89	14	10	17	21	62
70 - 79	6	5	12	16	39
60 - 69	21	5	8	12	46
50 - 59	17	10	6	8	41
40 - 49	10	12	10	7	39
30 - 39	18	15	6	5	44
20 - 29	13	18	8	3	42
10 - 19	15	15	6	5	41
1 - 9	4	17	7	1	29
(n_i)	125	112	102	105	(N) 444
(M_i)	50.0400	39.2857	61.4706	70.6190	(M) 54.8198
(s_i)	25.1595	27.0157	29.6267	23.9015	(SD) 28.9059

1. Ibid., p. 195.

On applying the correlation ratio technique, the following quantities are found:

$$\text{by (4)} \quad \eta^2 = 1 - \frac{310381.831}{370985.586} = .163359$$

$$\text{by (5)} \quad \eta^2 = \frac{60603.755}{370985.586} = .163359$$

$$\text{by (2) mean} \quad \eta^2 = \frac{4 - 1}{444 - 1} = .006772$$

$$\text{by (3) standard error } \eta^2 = \frac{1}{444 - 1}\sqrt{\frac{2(4 - 1)(444 - 4)}{444 + 1}} = .005498$$

$$\text{by (1)} \quad \frac{.163359 - .006772}{.005498} \quad \frac{.156587}{.005498} = 28.48$$

The resulting critical ratio, 28.48, indicates a significant correlation between the arrays or, in other words, the various means fluctuate significantly more than might be expected through chance. By applying Fisher's technique to the data of Table II, Institution B, and by substituting the proper values in (6), a z value of 1.67736 is obtained.

$$z = \frac{1}{2}\log_e\left\{\frac{(444 - 4)(.163359)}{(4 - 1)(1 - .163359)}\right\} = \frac{1}{2}\log_e 28.637515 = 1.677359$$

This value is sufficiently identical with the foregoing value that all calculations may be considered trustworthy.

The two methods clearly indicate that the data of Institution B show the average intelligence of the student body of this institution over a four-year period to have varied greatly from year to year. The different yearly groups cannot be thought of as being drawn from the same population; consequently, the total average of 54.8198 and the standard deviation of 28.9059 cannot be taken as representative of the student bodies of this institution.

If the data presented for Institution A were to be used to establish a standard of admission, it could be seen in the average of all the years, 64.3864, since the present tests demonstrated this average to be representative of all four years. The variation of the averages with respect to the total average can be found by the formula:

Variance	Degrees of Freedom	Sum of Squares	Mean Square	SD	Log_e SD
From Variation of Means	3	136.49494	45.49831	6.74524	1.908837
From Variation about Means	440	699.05818	1.58876	1.26046	.231477
Total	443	835.55312	1.88612	1.37336	1.677360 (z)

Relating z to its standard error by formula (8),

$$SE_z = \sqrt{\frac{1}{2}\,\frac{1}{3} + \frac{1}{440}} = .4096$$

a critical ratio of $\frac{1.677360}{.4096} = 4.10$ is obtained, which without doubt signifies that the variation of the means is statistically larger than that about the means. Again, for purposes of accuracy, the z obtained by Fisher's method can be checked by (7):

$$\text{(9) standard error } m_i = \frac{S_1^k\left\{n_i(m_i - M)^2\right\}}{N(k - 1)}$$

This quality (identical with the SD from the variation of the means in Fisher's method) is found to be .7148 for Institution A. The question must be answered as to what extent the yearly averages of Institution A may deviate from the total mean in order still to be within a range of chance fluctuation; or, in other words, when does the average of a

particular year become "off average", so that the college may say either the average ability of its student body has definitely improved over the previous year, or that it has significantly fallen off. Such a critical range, within which the yearly means may fluctuate, may be variously defined, such as ± 3 SE, or ± 5PE.[1] For the present illustration, a value of 3 SE in any one direction from the total mean may be considered as an "off mean." Since SE of the total mean for Institution A is .7148, the critical range can be defined as 64.3864 ± 3(.7148) or from 62.242 to 66.5308. Should any yearly mean fall outside this range, it would deviate significantly from the total mean, which is representative of the average ability of the student body of Institution A.

The same inquiry may be extended and the question be raised as to the limits between which any individual should fall in order that the institution may maintain the same characteristics of the student body as indicated by the total distribution. To answer this question for the present percentile data, in terms of the mean of the total distribution plus and minus a multiple of the standard deviation, is unsatisfactory due to the peculiar characteristics of the standard deviation of a percentile distribution. The best that can be done is to define a critical range in terms of the lowest and higher percentile rank on the national basis and keep the selection of students within that range. The quota of students within any one tenth could be made to approximate the percentages found in the total distribution as given in Table I. In case raw scores are available, the critical range for the selection of individual students is simply given by the total mean plus and minus a certain multiple of the standard deviation of the total distribution. Any student whose individual measure would fall outside this range could not be considered a good selection for Institution A.

While the establishment of the latter two critical ranges has been perfectly permissible for Institution A, it is not permissible for Institution B, because the analysis has shown that the total average is not representative of the yearly averages. In other words, Institution B did not have a definite standard of scholastic ability for its students over the four-year period studied.

One way of arriving at a representative average for Institution B is by combining[2] the set of four means, having due regard for their individual standard errors and weights. The process is as follows: The SE's of the four means are computed by the formula:

$$(11)\quad \text{standard error } m_i = \frac{s_i}{\sqrt{n_i}}$$

The weight w_i for each mean is computed by the formula:

$$(12)\quad w_i = \frac{1}{sd_i^2}$$

Each mean is then weighted by its w_i, and an average mean for the set of four is found by the formula:

$$(13)\quad M_{combined} = \frac{S_1^4 m_i w_i}{S_1^4 w_i}$$

The standard error of this $M_{combined}$ is found by the formula:

$$(14)\quad \text{standard error } M_{combined} = \frac{1}{\sqrt{S_1^4 w_i}}$$

The results of these computations are given in Table III.

TABLE III

THE COMBINATION OF THE FOUR HETEROGENEOUS MEANS FOR INSTITUTION B

m_i	sd m_i	w_i	$m_i w_i$
50.0400	2.2503	.1975	9.8829
39.2857	2.5527	.1535	6.0304
61.4706	2.9335	.1162	7.1429
70.6190	2.3325	.1838	12.9798
Total		.6510	36.0360

From this table, it is found that

$$M_{combined} = \frac{36.0360}{.6510} = 55.3548$$

1. J. B. Scarborough, Numerical Mathematical Analysis, p. 346. Baltimore: The John Hopkins Press, 1930.
2. Ibid., p. 328.

$$\text{standard error } M_{combined} = \frac{1}{\sqrt{.6510}}$$

$$= \frac{1}{.806846} = 1.2394$$

A representative standard for Institution B, therefore, is the $M_{combined}$, or 55.3548. A critical range could be set up by adding and subtracting 3 SE $M_{combined}$; this range would be between 51.6366 and 59.0730. Should Institution B be willing to set up any admission standard, having due regard to the average ability of its student body over the four-year period studied, the $M_{combined} \pm 3\ SD_{combined}$ would be the best standard under the circumstances. This, of course, does not mean that any institution may not arbitrarily set any other standard and critical range for the selection of its student material. Such arbitrary fixation, however, always implies the danger that the institution may be out of harmony with the standard it may have, if it draws without special effort on its natural resources. In other words, Institution B, in the long run, would be better off if it chose $M_{combined} \pm 3\ SD_{combined}$ as its standard and critical range, rather than the standard set, perhaps by the fourth year. In the former case, a certain continuity with respect to geographic and social-economic factors is preserved --factors which may not be consistent with any arbitrary standard.

Up to this point it has been illustrated how an institution may ascertain certain characteristics of its student body and how a quantitative expression of such characteristics may lead to the establishment of a definite standard and a critical range of selective admission. The same technique, however, may also be applied by an institution attempting to evaluate another institution, particularly a high school, with regard to the characteristics of the students admitted from that school. The question may be raised as to whether the high school has sent to the college homogeneous or heterogeneous material with regard to a certain measure.

To illustrate the two foregoing methods of analysis, two institutions are considered. In the first case, the high school has furnished the institution heterogeneous material, and in the second, homogeneous material. The data are given in Tables IV and V. By substituting in formula (1), as described on page 15, it is found that the critical ratio:

$$\lambda = \frac{.2936 - .0333}{.0230} = 11.32$$

and using formula (6), the value of z is found to be, z = 1.2447. By relating z to SE_z, .3596, a ratio of 3.46 is obtained, so that both methods clearly indicate the fluctuation of the means to be significantly

TABLE IV

DISTRIBUTIONS OF INTELLIGENCE PERCENTILE RANKS OVER PERIOD OF FIVE YEARS OF STUDENTS OF HIGH SCHOOL A ADMITTED TO HIGHER INSTITUTION

National Percentile Ranks	Distribution of Percentile Ranks					Total Distribution All Years
	1st Year	2nd Year	3rd Year	4th Year	5th Year	
90 - 100	8	-	4	-	5	17
80 - 89	4	-	6	-	2	12
70 - 79	5	1	1	3	1	11
60 - 69	3	1	2	5	1	12
50 - 59	3	3	3	4	2	15
40 - 49	3	4	3	1	1	12
30 - 39	-	5	5	-	1	11
20 - 29	-	4	2	-	3	9
10 - 19	-	5	5	5	1	16
1 - 9	-	4	-	2	-	6
(n_1)	26	27	31	20	17	(N) 121
(m_1)	75.77	31.67	54.68	45.00	62.65	(M) 53.60
(s_1)	17.30	19.05	28.11	25.30	28.61	(SD) 28.41

greater than might be expected through mere chance. The complete method of Fisher is given below.

Variance	Degrees of Freedom	Sum of Squares	Mean Square	SD	$\log_e$ SD
From Variation of Means	4	236.9743	59.2436	7.697	2.04083
From Variation about Means	116	570.1427	4.9150	2.217	.79615
Total	120	807.1170	6.7260	2.593	1.24468 (z)

TABLE V

DISTRIBUTION OF INTELLIGENCE PERCENTILE RANKS OVER PERIOD OF FIVE YEARS OF STUDENTS OF HIGH SCHOOL B ADMITTED TO HIGHER INSTITUTION

National Percentile Ranks	Distribution of Percentile Ranks					Total Distribution All Years
	1st Year	2nd Year	3rd Year	4th Year	5th Year	
90 - 100	9	8	8	6	3	34
80 - 89	7	5	4	5	8	29
70 - 79	8	3	4	5	4	24
60 - 69	4	4	3	3	-	14
50 - 59	4	6	1	2	2	15
40 - 49	4	4	-	4	4	16
30 - 39	2	2	5	-	-	9
20 - 29	-	2	-	1	-	3
10 - 19	1	1	-	1	-	3
1 - 9	1	-	-	1	-	2
(n_i)	40	35	25	28	21	(N) 149
(m_i)	70.00	66.14	73.00	67.86	74.04	(M) 69.77
(s_i)	22.69	23.39	22.09	24.76	17.43	(SD) 22.69

By substituting in formulas (1) and (6), the following values for school B are obtained:

$$\lambda = \frac{.0158 - .0270}{.0187} = -.6$$

$$z = -.27497$$

By relating z to its SE_z, .3584, a ratio of -.77 is obtained. The complete method of Fisher for Institution B is as follows:

Variance	Degrees of Freedom	Sum of Squares	Mean Square	SD	$\log_e$ SD
From Variation of Means	4	8.1216	2.0304	1.425	.35417
From Variation about Means	144	506.7561	3.5191	1.876	.62914
Total	148	514.8777	3.4789	1.865	-.27497 (z)

Both tests clearly indicate that the variation of the means is less than the variation about the means. This signifies that high school B during the period studied has sent to the higher institution student material which has been homogeneous with regard to intelligence percentiles. The average over the five years approximates the 70th percentile rank.

The same statement cannot be made for high school A, since its student material has been very heterogeneous. In this case, the college admitting the students cannot rely on the high school with regard to the ability of the students applying for admission. Every student from high school A, therefore, must be considered individually with regard to his ability. Should his measure of ability fall within the critical range set up by the admission office, he may be admitted; otherwise not, especially if the college wishes to keep its student body fairly homogeneous with regard to the particular ability. To attempt a rating of such a high school is not possible without serious limitations. On the other hand, a high school which has supplied fairly homogeneous student material to the higher institution can be rated with other high schools of the same type. How such a rating can be established, and how finally a series of constants can be worked out adjusting various high schools for particular measures, such as intelligence of students, average marks, and students' class ranks, has been shown by the writer elsewhere.[1] Such adjustment methods are only worth while if the fact has been established that the measure on which the rating or adjustment scheme has been based is representative of a homogeneous set or sets of values.

With the present statistical technique, it is possible to examine all high schools which have furnished student material to a higher institution, so that finally such high schools can be grouped into two, or more, categories--high schools which have sent fairly homogeneous material and those which have sent heterogeneous material to a higher institution over a period of years. Students could be admitted from the first class of high schools without detailed individual consideration, while students from the second group could not be admitted with the same confidence. Students from the latter schools probably should be subjected to an individual analysis of their scholastic aptitude, their final admission being determined by their meeting the standard and critical range of the selecting institution.

The present statistical technique, however, need not be confined to so-called initiatory measures, but may well be applied to measures of the finished product, such as marks, college grade points, number of quarters or semesters in residence. These techniques also may be applied to such measures as intelligence of students graduating from an institution of higher learning, the results of which would indicate whether or not an institution over a period of years graduates a homogeneous or heterogeneous product. Such investigations could be utilized for several purposes: For example, an analysis of initiatory intelligence of the student body of an institution might show that the institution selected homogeneous material of the 60th percentile rank, and that, on graduating, the intelligence of the students centers around the 70th percentile rank. From these facts, an institution could conclude that an internal selective process takes place to the amount of ten percentiles at the level at which the institution is operating. This selective process may be natural or excessive. If a corresponding difference on a nation-wide basis were only five percentiles, the institution would be forced to conclude its internal selective process to be five percentiles in excess of that which is reasonably expected. In other words, the particular institution has set standards for its work for which it does not select corresponding student material. In terms of dollars, this might mean, under certain circumstances, an enormous loss to the institutions. In terms of educational efficiency, it might indicate a waste of effort and probably a less satisfactory internal adjustment. While there is, in all probability, some selective process taking place, any excess is a sign that the institution is not doing what it intends to do. An institution which can satisfactorily graduate a large

1. "Forecasting Success of Students under the New Plan at the University of Chicago." Unpublished Report to the Dean of Students and University Examiner, May 9, 1933. See also, "A Method of Adjusting School Marks and Class Ranks to Compensate for Variations in Standards of Marking", published in the Bulletin of the American Association of Collegiate Registrars, (October, 1933), 1-14.

percentage of the students it admits is certainly functioning more efficiently from an educational point of view than is the institution which eliminates fifty to eighty per cent of its students before graduation.

The foregoing techniques should be useful also in coordinating higher institutions within a state, or within even a larger region. Particularly is this true when through need of economy, it is desirable to combine or eliminate institutions. By these techniques it should be possible to determine the institutions having the strongest units of a specialized type of education, and certainly that institution, or unit, which would show the highest level of ability and homogeneity in its finished product, all other considerations--such as aims and objectives, location, and personnel--being constant, would be the most desirable to retain.

A few words may be said about the limitation of the techniques herein recommended for the study of institutional differences. One of the limitations has already been discussed, namely, that the measures studied should follow the law of normal distribution. It was pointed out that percentile ranks do not follow this law and are, therefore, less applicable for analysis. This is not so serious, however, for E. S. Pearson has shown[1] that the sampling distribution of η^2 is remarkably insensitive to changes in population form. This is only true, however, if the number of cases, n_1, for each year is identical or practically identical. If the n_1 vary markedly from year to year, the η^2 test is only valid when the distributions are normal. Both techniques are less applicable, however, should the number of cases be too small; that is, $n_1 + n_2 \leq 20$, or if n_1 and n_2 in the formula for the standard error of z are greatly different from one another, especially when n_1 is very small. In this case, the z cannot longer be related to its standard error; but Fisher's tables of z distributions[2] should be used for the interpretation. In these tables, the probability is given that z in either 5% or 1% of random samples would surpass a given quantity. Wherever the entries of the tables are not sufficient to cover n_1 and n_2, the 5% or 1% points value can be computed by the formulas:[3]

$$(15)\quad z = \frac{1.6449}{h - 1} - .7843\left(\frac{1}{n_1} - \frac{1}{n_2}\right) \text{ for the 5\% point, and}$$

$$(16)\quad z = \frac{2.3263}{h - 1} - 1.235\left(\frac{1}{n_1} - \frac{1}{n_2}\right) \text{ for the 1\% point}$$

The symbol h in these equations represents the harmonic mean of n_1 and n_2:

$$h = \frac{2n_1 n_2}{n_1 + n_2}$$

An illustration will serve to make the use of this test clear. The four z values which have been computed in this paper are compared with the corresponding z value at the 1% point computed by (16). The results are as follows:

z Values, observed	n_1	n_2	Z Value, theory 1% point calculated
Data, Institution A - .5897	3	436	.6358
Data, Institution B 1.6774	3	440	.6357
Data, High School A 1.2447	4	116	.5984
Data, High School B -.2750	4	144	.5930

The results reveal that by this test also the observed z values for Institution A and High School B are not significant.

Pearson gives a similar test[4] by defining

$$\lambda = \frac{\text{observed } \eta^2 - \text{mean } \eta^2}{\text{standard deviation } \eta^2}$$

and asking the question as to what the value of λ, or of η^2 (η^2 = mean η^2 + λ standard deviation η^2), must be in order that the

1. E. S. Pearson, op. cit., p. 118.
2. Fisher, op. cit., p. 212.
3. Ibid., p. 201.
4. K. Pearson, op. cit., Part II, p. xli.

tail cut off the distribution curve of η^2 may be equal to $\frac{1}{100}$ths or $\frac{2}{100}$ths of the total frequency. Thus λ_1 and λ_2 measure the deviations from the mean η^2, at which the chances of an η^2 occurring are respectively less than $P_1 = .01$, or $P_2 = .02$, supposing no correlation exists. When a value of η^2 gives a value of λ greater than λ_1, then the existence of correlation may be reasonably predicted. When η^2 gives a value less than λ_2, then the existence of any correlation remains doubtful. The various values of λ_1, λ_2, P_1, and P_2 are here reproduced:

η^2 Values, observed		Theoretical Values						Observed
		K	N	λ_1	P_1	λ_2	P_2	λ
Data, Institution A	.0021	4	440	3.20	.012	2.80	.020	-.85
Data, Institution B	.1634	4	444	3.20	.012	2.80	.020	28.48
Data, High School A	.2936	5	121	3.14	.011	2.68	.020	11.32
Data, High School B	.0158	5	149	3.14	.011	2.68	.020	-.6

Again, as this test shows, the data for Institution B and High School A indicate significant correlation; that is, the variation of the means in each case is larger than might be expected through mere chance.

With regard to the computation of the different intermediate quantities occurring in the two techniques, it may be said that the complete calculation of standard deviations for each individual year is not necessary, since η^2 can be obtained from sums involving these standard deviations squared. The same is true with regard to the quantity, "sum of squares from variation of means",

$$\frac{1}{N} S_1^k \left\{ n_1(m_1 - M)^2 \right\}$$

which can easily be obtained from the relation:

$$\frac{1}{N} S_1^k \left\{ n_1(m_1 - M)^2 \right\} + \frac{1}{N} S_1^k (n_1 s_1^2) = \frac{1}{N} S_1^N (x - M)^2$$

The use of the foregoing intermediate quantities and relations will reduce the labor of calculation an appreciable amount.

Conclusion

1. Two statistical methods have been presented by means of which an adequate study of institutional differences of student bodies in particular measures has been made possible. The first method was Pearson's technique of testing the significance of the correlation ratio η_{xy}, if y_i represents the group number or the number of years and x_i the particular measure which is studied. The second method was Fisher's procedure of "analysis of variance." This method consists of the calculation and examination of the significance of z, the difference of the natural logarithms of the estimate of variance from the variation of the means and the variance from the variation about the means. The two tests lead to essentially identical results, since z is a single-valued function of η^2.

2. It has been pointed out how the statistical study of institutional differences may be made the basis for: (a) the establishment of standards of admission, (b) the rating of high schools, (c) the coordination of institutions or units of an institution within a state or regional area, and finally how such a study may form a basic procedure in the determination of the educational efficiency of institutions.

THE VARIATION OF HIGH SCHOOL SENIOR AND COLLEGE FRESHMAN CLASSES[1]

by
Vergil E. Lyon

The purpose of this investigation is to determine the variation of the mean intelligence of high school senior and college freshman classes. The variation to be considered will be between schools and between successive classes in the same schools. The specific questions to be answered are: What is the variation of the mean intelligence of students graduated from different high schools? What is the variation of the mean intelligence of students admitted to different colleges? What is the relation of the mean intelligence of graduating classes to the size of the school? To what extent do classes graduated from the same school vary in intelligence? To what extent do classes admitted to the same colleges vary in intelligence?

The determination of the variation of high school senior classes will be of value in admitting students to institutions of higher learning. If high schools vary considerably in the ability of their graduates, colleges admitting these graduates should be made aware of this fact. Students coming from a high school with a very low mean intelligence could not be admitted with the same assurance as those coming from a high school with a high mean intelligence. The variation from year to year within the same school is also of significance in college admission. Unless high schools graduate students of a consistent quality no rating of the high school can be made. If high schools are consistent in graduating students of certain qualities, colleges can then adjust their entrance requirements to take care of differences between high schools. In the case of colleges, prospective students and high school principals should know the level of ability accepted in certain colleges. Colleges vary considerably, and guidance should be given to the students in this matter. Colleges should also know how consistently they are drawing students of a certain level of ability.

The data[2] involved in this study consist of the results of tests administered to high school seniors in Wisconsin high schools during 1929, 1930, 1931, 1932, and 1933. These tests were given by the high school principals and scored at the University of Wisconsin.[3] In 1929 the Ohio State University Psychological Test, Form 14 was administered to 14,967 high school seniors. In 1930, 1931, and 1932, the 1930, 1931, and 1932 Editions of the American Council Psychological Examination were administered to 16,329, 20,000, and 23,242 high school seniors. In 1933 the Henmon-Nelson Test of Mental Ability, High School Examination, Form A, was administered to 26,122 high school seniors. After these examinations were scored the scores were transmuted into percentile ranks and recorded. The fact that the results of these tests are recorded as percentile scores makes a very serious limitation to one part of this study, to which reference will be made later.

Since some reduction in the number of high schools to be considered was necessary only schools graduating at least thirty seniors in each of the five years for which

1. This study has been made possible by funds from the Wisconsin Alumni Research Foundation.
2. The writer is indebted to Professor V. A. C. Henmon for making these data available, and for many helpful suggestions.
3. A detailed description of this testing program is given in V. A. C. Henmon and F. O. Holt. "A Report on the Administration of Scholastic Aptitude Test to 54,000 High School Seniors in Wisconsin in 1929 and 1930." Bulletin of the University of Wisconsin, (June, 1931).

data were available were included in this study. The total number of schools involved was one hundred and eight and the total number of students was approximately 60,000.

The twenty institutions of higher learning in Wisconsin receive copies of these test results and then determine the mean percentile rank for their freshman classes. These means and the distribution of freshman scores constitute the data on which the study of the variability of college freshman classes will be made.

Since percentile scores on five different tests are to be compared it is necessary to study the equivalence of these scores. Since the percentile scores are based on such a large number of cases, at least 14,967, and since the senior high school population probably does not vary over a five-year period the percentile scores are comparable.

The question of the extent to which these different tests measure the same ability can only be partially answered with the available data. The correlation between the Henmon-Nelson Test and the American Council Psychological Examination; 1929, 1930, and 1931 Editions; for 639 university freshmen is .73.[1] This correlation when corrected for attenuation becomes .85. This correlation is sufficiently high to indicate that these gests measure the same ability. Further evidence on this point is given by the similarity of the content.

The first step in the computation consisted in calculating the mean percentile score of each school for each of the five years. The mean percentile score for the total group from each school without regard to the year was also found. In most cases all computations were made on the ungrouped data. The means for the one hundred and eight high school senior classes are given in Tables I and III. The means for the twenty colleges are given in Tables II and IV. Inspection of these tables reveals some very striking figures. The means of the individual school means are fairly constant from year to year but the individual school means within any given year vary widely. In 1929 the range of these means

TABLE I

THE MEAN INTELLIGENCE AND VARIABILITY FOR 108 HIGH SCHOOLS

	Mean Number	Mean Intelligence Percentile						Measures of Variability		
		1929	1930	1931	1932	1933	Total	Range	A.D.	z
1	52	55.9	47.4	56.2	59.1	49.0	53.4	11.7	4.3	.27
2	44	47.1	51.8	44.3	44.9	53.7	58.8	9.4	3.6	
3	149	49.7	56.7	48.3	39.2	48.8	46.3	10.5	3.1	.68
4	265	47.7	49.4	48.9	44.3	51.9	48.3	7.6	2.0	.51
5	130	44.1	39.8	41.4	44.3	34.7	40.9	9.6	2.9	
6	72	63.0	65.0	74.1	69.0	67.6	67.9	11.1	3.1	.39
7	65	52.5	45.8	47.2	47.6	43.0	46.8	9.5	2.3	
8	79	49.1	42.7	55.8	45.6	44.3	47.2	13.1	3.9	.51
9	201	60.5	53.1	60.9	53.4	51.9	55.6	9.0	3.7	.71
10	64	36.1	44.9	44.1	45.8	54.0	44.6	17.9	4.0	.53
11	44	60.8	66.5	40.2	49.3	46.5	51.3	26.3	8.5	.91
12	54	55.7	49.0	44.2	44.3	49.0	48.1	11.5	3.4	
13	104	42.1	46.0	47.3	47.8	44.2	45.7	5.7	1.8	-.17
14	50	48.2	52.9	48.3	55.0	49.2	50.7	6.8	2.6	
15	55	42.1	55.8	46.6	41.2	55.7	48.6	14.6	6.0	.67
16	48	41.7	44.6	47.0	49.7	45.1	45.7	8.0	2.2	-.38
17	42	52.6	43.3	47.2	43.2	43.3	45.7	9.4	3.1	
18	37	46.7	60.3	66.6	47.6	55.7	55.5	19.9	6.6	.59
19	36	44.2	53.1	61.1	54.6	53.5	52.9	16.9	3.9	.25
20	251	47.7	48.1	54.5	60.3	49.0	52.4	12.6	4.5	1.12
21	52	51.4	42.8	44.8	45.2	48.6	46.2	8.6	2.7	

1. Data furnished by Professor V. A. C. Henmon.

TABLE I (Continued)

	Mean Number	Mean Intelligence Percentile						Measures of Variability		
		1929	1930	1931	1932	1933	Total	Range	A.D.	z
22	43	48.5	49.7	48.3	36.4	59.5	47.5	23.1	5.4	.71
23	48	36.1	37.9	55.5	45.2	60.4	46.9	24.3	8.7	1.08
24	48	45.1	46.3	45.2	44.5	52.3	46.9	7.8	2.4	-.19
25	36	38.7	47.2	50.9	53.4	47.3	45.5	16.2	6.0	.48
26	242	47.4	37.1	43.9	48.1	46.7	45.2	11.0	3.3	.84
27	191	52.6	43.0	40.5	45.0	48.2	46.1	12.1	3.7	.83
28	166	42.8	40.0	42.2	44.6	53.6	45.0	13.6	3.8	.91
29	56	57.6	76.7	61.1	68.3	43.2	59.8	33.5	9.1	1.30
30	43	66.2	69.6	58.1	54.3	41.0	45.5	28.6	8.2	1.05
31	45	44.9	45.9	44.3	51.7	49.5	47.6	7.4	2.7	
32	65	26.9	37.3	35.7	37.8	42.9	36.7	16.0	3.7	.52
33	52	45.8	48.3	46.8	45.5	43.9	45.9	4.4	1.2	-.89
34	77	56.9	61.0	58.0	51.1	58.0	56.8	9.9	2.5	
35	297	66.3	60.1	71.9	66.8	55.2	63.7	16.7	5.2	1.57
36	52	51.8	39.4	40.1	38.3	48.7	43.2	13.5	5.2	.44
37	92	40.2	51.3	40.4	47.0	51.7	45.9	11.5	4.7	.66
38	78	46.4	43.0	40.0	44.6	48.4	44.3	8.4	2.4	
39	52	51.1	44.3	47.6	36.8	41.6	44.2	14.3	4.1	
40	267	57.6	58.0	55.8	53.7	49.2	55.4	8.8	2.6	
41	209	54.2	48.3	55.0	54.5	48.5	52.0	6.7	3.0	
42	57	42.4	53.4	54.7	57.6	48.2	51.5	15.2	4.7	.43
43	68	68.7	69.1	80.0	63.1	61.5	68.2	18.5	5.0	.70
44	238	55.2	60.2	53.2	49.7	55.0	54.4	10.5	2.6	
45	119	51.0	44.3	44.4	45.5	50.0	47.3	6.7	2.8	
46	95	41.9	44.2	47.6	47.4	50.6	46.5	8.7	2.6	
47	53	42.6	41.9	35.4	41.0	35.8	39.4	7.2	3.0	
48	48	55.2	43.1	43.7	51.3	43.5	46.8	12.1	4.6	
49	62	44.4	36.1	48.4	46.1	47.3	44.4	12.3	3.4	
50	108	49.7	47.4	51.1	39.8	40.6	45.1	11.3	4.5	.61
51	99	57.5	44.4	46.0	51.8	55.9	51.5	13.1	4.7	.73
52	45	44.7	44.8	49.1	46.5	39.1	44.8	10.1	2.4	
53	42	44.7	30.7	37.8	37.0	39.8	38.1	14.	3.4	
54	88	54.2	52.9	45.5	48.9	42.5	48.9	11.7	3.8	.51
55	104	51.6	51.5	48.4	49.3	43.5	48.5	8.1	2.4	
56	60	47.4	39.8	36.1	38.5	48.1	41.8	12.0	4.6	.40
57	71	43.4	61.4	50.0	44.6	36.7	48.4	24.7	7.0	1.08
58	78	34.2	34.5	32.3	34.2	41.0	35.3	8.7	2.3	
59	58	43.4	30.6	41.0	44.6	41.7	40.5	14.0	3.7	.32
60	38	32.3	43.5	40.3	53.6	38.0	42.5	21.3	5.8	.67
61	290	44.4	50.4	48.1	46.6	51.4	48.5	7.0	2.2	
62	45	54.4	44.1	42.8	47.6	33.7	44.2	20.7	5.1	.55
63	84	52.8	43.3	43.6	46.5	48.0	46.8	9.5	2.8	
64	65	56.6	53.0	45.1	42.4	45.8	47.6	14.2	4.8	.47
65	94	60.6	42.6	48.8	48.3	48.5	49.6	18.0	4.2	.76
66	45	51.0	49.6	45.0	50.9	53.7	50.7	8.7	2.1	
67	46	52.1	41.3	46.8	44.3	43.9	45.5	10.8	3.0	
68	70	61.3	64.6	59.9	42.0	55.9	58.6	12.6	3.9	.35
69	259	57.9	59.2	57.3	35.8	55.4	56.8	3.8	1.3	
70	95	49.8	49.7	45.0	46.2	47.3	47.4	4.8	1.7	
71	82	62.9	54.4	54.6	50.7	53.9	55.2	12.2	3.0	
72	89	40.9	38.4	41.9	46.3	49.0	43.5	10.6	3.5	
73	57	50.3	43.0	38.7	39.1	43.3	42.6	11.6	3.2	
74	58	44.1	38.2	40.2	37.3	48.2	41.9	10.9	3.7	
75	99	53.0	42.1	46.2	37.3	43.5	43.5	15.7	4.0	.69
76	281	49.2	57.4	54.0	55.4	55.5	54.5	8.2	2.1	
77	164	62.2	64.4	60.1	59.6	63.7	62.0	4.8	1.7	
78	69	51.1	47.7	54.2	45.7	49.0	49.5	8.5	2.5	
79	83	57.5	53.3	52.5	59.6	54.7	56.3	7.1	2.6	
80	50	54.0	36.9	47.2	28.4	35.8	39.6	25.6	7.9	.95

TABLE I (Continued)

	Mean Number	Mean Intelligence Percentile						Measures of Variability		
		1929	1930	1931	1932	1933	Total	Range	A.D.	z
81	54	36.7	31.0	33.9	37.6	35.9	33.1	9.1	3.0	
82	180	42.8	40.3	35.4	39.9	40.6	39.7	7.4	1.8	
83	96	46.4	50.2	46.7	40.4	48.7	46.4	9.8	2.5	
84	90	53.9	50.3	44.4	42.2	45.3	46.4	11.7	3.7	.42
85	314	50.0	56.6	55.5	47.9	46.3	50.3	10.3	3.6	.95
86	86	46.3	38.4	42.7	39.8	41.0	41.6	7.9	2.3	
87	84	41.0	50.2	46.3	48.3	45.5	46.4	9.2	2.4	
88	50	40.8	37.2	34.3	45.7	42.0	40.4	11.4	3.3	
89	91	39.0	42.5	34.0	38.0	45.4	40.3	11.4	3.4	
90	71	39.2	43.7	49.9	45.3	48.1	45.4	10.7	3.0	
91	86	73.4	51.9	52.1	54.4	47.2	54.6	326.2	6.3	1.19
92	184	54.3	50.7	53.4	48.7	61.2	54.0	12.5	3.3	
93	58	55.5	47.6	52.8	50.3	51.2	51.3	7.9	2.1	
94	43	44.1	53.0	58.4	42.7	53.3	50.1	15.7	5.6	.47
95	203	49.4	47.3	53.6	47.1	52.7	50.1	6.5	2.5	
96	175	61.2	58.8	65.0	65.1	61.4	62.4	6.3	2.2	
97	207	54.9	51.4	55.6	54.6	57.2	55.0	5.8	1.4	
98	64	71.1	62.5	58.0	55.9	43.4	55.9	27.7	7.3	1.15
99	44	47.4	49.4	49.1	53.4	42.0	47.9	11.4	2.9	
100	136	54.5	51.3	49.3	54.4	55.0	52.9	5.7	2.1	
101	328	59.4	57.2	60.4	59.1	58.0	58.8	3.2	1.0	
102	147	57.8	64.1	55.2	58.7	59.8	58.8	8.9	2.2	
103	167	60.0	57.9	45.2	45.7	42.5	48.7	17.5	6.6	1.24
104	423	61.5	54.8	55.9	58.3	54.4	56.8	7.1	2.3	
105	317	62.7	67.1	63.4	68.6	63.9	65.4	5.9	1.6	
106	68	55.6	51.9	57.7	78.4	63.6	61.8	26.5	7.8	1.13
107	589	55.5	66.3	66.1	65.0	65.2	63.7	10.8	3.2	1.44
108	363	65.4	60.6	60.3	59.2	54.5	59.7	10.9	2.6	1.02

TABLE II

THE MEAN INTELLIGENCE AND VARIABILITY FOR 20 COLLEGES

	Mean Number	Mean Intelligence Percentile						Measures of Variability		
		1929	1930	1931	1932	1933	Total	Range	A.D.	z
1	77	69.9		73.3	75.6	72.7	72.9	5.7	1.6	-.08
2	1550	67.3	69.8	73.6	70.4	71.7	70.3	6.3	1.7	1.44
3	36	65.8		81.0	69.8	78.1	73.4	15.2	5.9	.46
4	141	65.8		69.5	72.5	72.5	69.8	6.7	2.4	.46
5	22	65.0		52.5	59.0	60.7	58.9	12.5	3.6	-.06
6	260	61.2		69.6	71.6	68.2	68.2	10.4	3.0	.98
7	88	60.1	62.4	56.4	68.2	67.3	63.1	11.8	3.7	.59
8	305	59.2		66.0	74.9	74.7	69.1	15.7	6.1	1.66
9	28	57.7	53.3	59.4	62.6	48.3	56.9	14.3	4.2	-.04
10	67	56.1		57.8	59.6	69.0	60.1	12.9	3.9	.44
11	187	54.7		57.5	64.1	52.3	57.4	11.8	3.6	.99
12	163	53.7		53.9	55.6	58.4	55.4	4.7	1.6	-.09
13	186	53.2		55.7	60.7	59.1	57.6	7.5	2.7	.46
14	171	51.5		57.4	62.3	62.0	59.1	10.8	3.8	.74
15	210	49.6		52.9	51.7	53.5	52.2	3.9	1.3	-.38
16	150	49.4		50.2	52.6	56.1	52.1	6.7	2.3	.29
17	146	49.3	50.5	52.1	56.4	53.3	52.4	7.1	2.1	.17
18	80	47.8		56.8	57.5	58.6	55.5	10.8	3.5	.42
19	186	45.8		50.3	60.5	61.7	55.0	15.9	6.5	1.53
20	202		70.4	73.7	73.4	74.2	73.0	3.8	1.2	.00

TABLE III

THE DISTRIBUTION OF MEAN PERCENTILE SCORES FOR 108 HIGH SCHOOLS

Percentile scores	1929	1930	1931	1932	1933	Total
80 - 84			1			
75 - 79		1		1		
70 - 74	2		2			
65 - 69	4	6	3	5	2	3
60 - 64	11	10	7	3	7	5
55 - 59	16	8	16	12	14	16
50 - 54	23	22	14	19	19	18
45 - 49	20	19	32	34	33	41
40 - 44	23	25	23	17	24	18
35 - 39	6	13	6	14	7	6
30 - 34	2	4	4	1	2	1
25 - 29	1			2		
Number	108	108	108	108	108	108
Range	26.9 - 73.4	30.6 - 76.7	32.3 - 80.0	27.6 - 78.4	33.7 - 67.6	33.1 - 68.2
Mean	50.69	49.12	49.77	48.52	48.98	49.54
Sigma	8.83	9.35	8.98	8.60	7.30	6.94

TABLE IV

THE DISTRIBUTION OF MEAN PERCENTILE SCORES FOR 20 COLLEGES

Percentile scores	1929	1930	1931	1932	1933	Total
80 - 84			1			
75 - 79				1	1	
70 - 74		1	3	5	5	4
65 - 69	5	1	3	2	3	3
60 - 64	2	1		5	3	2
55 - 59	3		7	5	4	8
50 - 54	4	2	6	2	3	3
45 - 59	5				1	
Number	19	5	20	20	20	20
Range	45.8 - 69.9		50.2 - 81.0	51.7 - 75.6	48.3 - 78.1	52.1 - 73.4
Mean	57.00[1]	61.28	60.98	63.95	63.47	61.62
Sigma	6.78		9.07	7.38	8.35	7.40

was from 73.4 to 26.9. The corresponding range for the colleges is from 69.9 to 45.8. Similar ranges are found in the other years. There are a few high schools in the state in which approximately half the graduating class is in the lower fourth of the total high school senior class population. On the other hand there are some high schools in which approximately one half the graduating class is in the upper fourth of the total population. Such differences indicate the need of recognizing differences between high schools and colleges as well as between individuals. Certainly the curriculum and methods of teaching cannot be the same in two schools graduating classes with means of 26.9 and 73.4 nor can students from these two schools be accepted as equally promising college material. These differences are so large that ranks in graduating class and probably average high school grades can mean little unless the mean for the class is known. A student may be at the top of the class in one school while if he had gone to another school he might have

1. All means and sigma given in this table were computed from ungrouped data.

occupied an entirely different position. The differences between colleges are not as large as between high schools because the selection of college students is even more rigorous than high school graduation. However, there is considerable variation in this respect since some colleges admit students having the same percentile score as is found in the total high school senior class and some others in which the mean is at or above the seventy-fifth percentile point. The quality of work which can be expected of students in such colleges cannot be the same.

In connection with these differences one might ask if size of graduating class or of school bears any relation to level of intelligence. Henmon and Holt[1] found for the 1929 data a statistically significant difference in favor of schools graduating forty or more seniors as compared to those schools graduating between 10 and 39 seniors. Skewes and Wilson[2] found no correlation between mean percentile score and size of class using data from 1929, 1930, and 1931 and an unselected group of one hundred high schools.

Since only schools graduating at least thirty seniors per year have been considered in this study direct comparisons with these two studies cannot be made. The correlation found between the means for the five years combined, total means in Table I, and the mean size of class for the one hundred and eight schools used in this study is .44. This is large enough to have some significance. The scatter-diagram from which this correlation was computed showed many schools with less than the average number in graduating class to be above the average in mean percentile score but few schools of above average size which were below the average in mean percentile score.

The second major problem of this study is the determination of the variation between classes within schools. To what extent do classes graduated from the same school vary? To what extent do classes admitted to the same college vary? Few studies have been made of this type of variation due to the lack of data for successive years or classes and the lack of an adequate statistical technique for the analysis of such variation. The usual technique for determining the significance of differences is inadequate since in this problem we are testing the significance of a series of differences rather than one difference.

One measure of this variability is the correlation between means of successive classes. Several studies using data from Wisconsin high schools have used this measure. Skewes and Wilson found correlations of .27, .24, and .40 between the mean percentile scores for 1929 and 1930, 1929 and 1931, and 1930 and 1931, respectively. Table V gives the correlations between successive medians for all Wisconsin high schools. Tables VI and VII give the correlations of medians of schools graduating less than forty seniors and forty or more seniors.

TABLE V[3]

THE CORRELATION BETWEEN MEDIANS OF SUCCESSIVE SENIOR CLASSES, ALL SCHOOLS

	1930	1931	1932
1929	.363 (434)[4]	.303 (427)	.291 (438)
1930		.390 (438)	.363 (448)
1931			.452 (447)

TABLE VI

THE CORRELATION BETWEEN MEDIANS OF SUCCESSIVE SENIOR CLASSES, SCHOOL WITH LESS THAN 40 SENIORS

	1930	1931	1932
1929	.318 (319)	.232 (313)	.221 (321)
1930		.332 (322)	.302 (331)
1931			.393 (326)

TABLE VII

THE CORRELATION BETWEEN MEDIANS OF SUCCESSIVE SENIOR CLASSES, SCHOOLS WITH FORTY OR MORE SENIORS

	1930	1931	1932
1929	.582 (113)	.600 (113)	.604 (114)
1930		.731 (116)	.662 (116)
1931			.711 (120)

1. Op. cit., p. 26.
2. George Skewes and G. Alvin Wilson, "A Study of the Consistency of High Schools in Graduating Pupils Adapted for College Work", unpublished material.
3. Data for Tables V, VI, and VII were furnished by Professor V. A. C. Henmon.
4. Figures in parentheses indicate the number of schools on which the correlation is based.

The correlations for the one hundred and eight schools involved in this study are given in Table VIII. The correlations for the high schools are consistently higher than those found by Skewes and Wilson. This is probably due to the difference in size of schools. Skewes and Wilson used an unselected group of one hundred schools containing some schools with as few as five seniors.

The correlations for the twenty Wisconsin colleges are given in Table IX. Since the data for all colleges was not collected in 1930 this year is omitted from these calculations. These correlations are slightly higher than the corresponding correlations for the high schools and indicate that colleges admit classes of more nearly the same ability from year to year than high schools graduate.

TABLE VIII

THE CORRELATION BETWEEN MEAN INTELLIGENCE PERCENTILES FOR 108 HIGH SCHOOLS

	1930	1931	1932	1933
1929	.609	.597	.543	.341
1930		.733	.669	.504
1931			.736	.631
1932				.579

TABLE IX

THE CORRELATION BETWEEN MEAN INTELLIGENCE PERCENTILES FOR 20 WISCONSIN COLLEGES

	1931	1932	1933
1929	.79	.77	.68
1931		.84	.79
1932			.78

While these correlations between names of successive classes give some indication of the variability from year to year a measure giving a more precise statement is needed. Three additional measures have been used in this study. These measures are: (1) the largest difference or the range of the yearly means, (2) the average absolute deviation of the yearly means around the total means, and (3) Fisher's z function which is one half the difference between the natural logarithms of an estimate of the variance between classes and an estimate of the variance within classes.

The largest differences between means are presented in Tables I and II under the heading "Range." Table X gives these differences in a frequency distribution. The range of these differences is from 3.2 to 33.5 for the high schools and from 3.8 to 15.9 for the colleges. The range of differences is much less for the colleges and the mean difference is also less for colleges. This confirms the results of the correlations between successive classes. Colleges select more nearly the same type of classes from year to year than high schools graduate. It is difficult to establish a point above which these differences are significant but it would probably not be larger than twenty. Classes which varied in mean percentile score by twenty points would appear to be significantly different.

TABLE X

THE DISTRIBUTION OF LARGEST DIFFERENCES FOR 108 HIGH SCHOOLS AND 20 COLLEGES

Differences	High Schools	Colleges
33 - 35	1	
30 - 32		
27 - 29	2	
24 - 26	6	
21 - 23	2	
18 - 20	4	
15 - 17	9	3
12 - 14	17	3
9 - 11	32	5
6 - 8	26	5
3 - 5	9	4
Number	108	20
Mean	12.21[1]	9.72[1]
Sigma	5.91[1]	3.91[1]

The second measure of variation from year to year is the average absolute deviation of the yearly means around the total mean. These deviations are given in Tables I and II under the heading "A.D." They are given in a frequency distribution in Table XI. These tables reveal a wide range in average absolute deviation. The most variable high school has an average absolute deviation of 9.1 while the least variable has an average absolute deviation of only 1.0. The corresponding figures for the colleges are 6.5 and 1.2. The mean average deviation for the high schools is 3.74 and more than one half the schools fall

1. Computed from ungrouped data.

between 2.0 and 4.0. Comparison of these deviations for high school senior classes and college freshman classes indicates that college freshman classes are less variable.

TABLE XI

THE DISTRIBUTION OF AVERAGE ABSOLUTE DEVIATIONS FOR 108 HIGH SCHOOLS AND 20 COLLEGES

Average absolute deviations	High Schools	Colleges
9.0 - 9.4	1	
8.5 - 8.9	2	
8.0 - 8.4	1	
7.5 - 7.9	2	
7.0 - 7.4	2	
6.5 - 6.9	2	
6.0 - 6.4	3	1
5.5 - 5.9	2	2
5.0 - 5.4	5	
4.5 - 4.9	8	
4.0 - 4.4	5	1
3.5 - 3.9	14	6
3.0 - 3.4	18	1
2.5 - 2.9	16	1
2.0 - 2.4	18	3
1.5 - 1.9	5	3
1.0 - 1.4	4	2
Number	108	20
Mean	3.75	3.23[1]
Sigma	1.74	1.53[1]

It is very difficult to interpret this measure of variability except in comparing one school with another. An average absolute deviation of a given size may be due to this deviation in each class or to a very large deviation in one or two classes. Practically the second situation would be the most significant.

As one would expect the average absolute deviation is related to the size of the classes. For the high school data the correlation between these two measures is -.45. Inspection of the scatter-diagram on which this correlation was computed suggested the possibility of a non-linear relation. However, since the correlation ratio was only .47 the product-moment correlation can be used to represent this relationship. This correlation would probably have been even higher had smaller schools been included.

The question of the relation between the average absolute deviation and the total mean was also found. This correlation is .08. There is no tendency for schools on one level to vary more from year to year than others.

The third measure of variability employed is derived from an estimate of the variance within years and an estimate of the variance between years. This method for the analysis of variance into two or more parts was developed by Fisher[2] and its application to this particular has been suggested by Reitz.[3]

This technique tests the significance of the difference between the part of the total variance which is due to variation of individuals within each successive class and the part due to variation in the means of successive classes. In case the part of the total variance contributed by the variation between classes is significantly larger than that part due to the variation within classes the successive classes cannot be considered as coming from the same population. Or, stated in different terms, in such a case the system of means could not be considered as arising by drawing successive samples from the same parent population. Some factor other than chance must have been operating. The value whose significance is tested is not the difference between these two estimates of the variance but one half the difference between the natural logarithms of these estimates. This quantity is known as (z). In case the number of degrees of freedom corresponding to the larger variance, (n_1), and the number of degrees of freedom corresponding to the smaller variance, (n_2), are large or moderate in size and nearly equal the significance of (z) can be determined by relating it to its standard deviation. Since in the data for the high schools one of the degrees of freedom will be four and the other will be not less than 180, this test

1. Computed from ungrouped data.
2. R. A. Fisher, Statistical Methods for Research Workers. Oliver and Boyd: Edinburgh, (1930), Third Edition, pp. 178-215.
3. Wilhelm Reitz, "Statistical Techniques for the Study of Institutional Differences", Journal of Experimental Education, III (September, 1934), pp.

of significance cannot be used. Fisher gives tables[1] which permit an interpretation of the significance of (z) in terms of probability. These tables give (z's) corresponding to probabilities of .05 and .01 for certain values of n_1 and n_2. The z corresponding to a probability of .01, n_1 of 4, and n_2 of 60 is .6472. The corresponding z for an n_1 of 4 and n_2 infinitely large is .5999. The number of degrees of freedom corresponding to the smaller variance, n_2, involved in the data of this study do not occur in these tables. For this reason some arbitrary limit must be set since it would be impractical to complete these tables for all the different n_2's involved. The point z = .6472 has been arbitrarily chosen. The probability of a z of this size arising through chance is not far from .01. In terms of the problem of this study a school in which (z) equals or is greater than .6472 cannot be considered as drawing its senior classes from the same parent population. Or in the usual terms, the differences between the successive means are statistically significant.

Before presenting the data on this point it is necessary to consider a limitation of the application of this technique to these data. This technique assumes a normal distribution of the variate in the population from which the samples are drawn. The variate studied in this investigation has been measured in percentile units which yield a rectangular distribution rather than a normal distribution. This fact undoubtedly has affected the results probably in the direction of exaggerating the differences. The results must be accepted as tentative until this test of significance has been revised to take care of rectangular distributions.

The z's for the most variable high schools are given in Table I and for the colleges in Table II. The z's have been computed for those schools in which the variation seemed by inspection to be significant and a few others. If one accepts the point .6472 as the limit above which differences between successive means are significant twenty-eight of the one hundred and eight schools are more variable from year to year than would be expected from chance. In these twenty-eight schools successive classes cannot be considered as drawn from the same population nor can the total mean be used as a measure of the ability of students from these high schools. There are a few negative values of z given in Tables I and II. This indicates that the estimate of the variance between classes is less than the estimate of the variance within classes. None of these negative values are larger than would be expected through chance.

There are six of the twenty colleges which have z's greater than .6472. These freshman classes cannot be considered as drawn from the same population.

A word of caution should be given in interpreting these results. This technique gives only an indication of the statistical significance of certain values of z. The probability of getting a z of .6472 or larger through chance is less than .01. This does not tell one much about the practical significance of the variation. School Number 108 has a z value of 1.02 and yet the variations from year to year are relatively small and probably not of any practical importance. School Number 10 has a z of only .53 and yet its variation from year to year is larger when measured both by range or largest difference and average absolute deviation. There is no reason why a certain deviation should be considered practically significant in a large school but not in a small school. The conclusion one can reach regarding these twenty-eight high schools and six colleges is that the means of successive classes vary so widely in comparison to the variability within classes that some factor or factors other than chance have been operating.

Earlier in this study it was shown that the variation in means of high schools is very large. In 1929 these means varied between 26.9 and 73.4 and the means for the total five-year period varied between 33.1 and 68.2. It was pointed out that ranks in graduating classes need some adjustment to take account of this variation. At the University of Wisconsin predictions of

1. Op. cit., pp. 212-215.

university success are made on the basis of rank in graduating class and the percentile scores used in this study. The middle student coming from high school number 31 which has a total mean percentile score of 33.1 would have the same rank in graduating class as the middle student coming from high school number 43 which has a total mean percentile score of 68.2. These students cannot be equally promising college material. Reitz[1] has suggested and applied a technique for correcting ranks in graduating classes and also mean high school grades for this variation between schools. He points out, however, that such a technique is dependent on the variability from year to year within the same school. Since the z's in twenty-eight of the schools employed in this study are so large that the total means cannot represent these schools the adjustment could not be applied to them.

If such an adjustment of ranks is needed one would expect to find that the predictions of university success would be too high for students coming from high schools with low total means and too low for students from high schools with high total means. Table XII indicates that this is the actual situation. In every year the predictions of university success in terms of grade point averages are too low for the students coming from the thirty high schools with the highest total means. In every year but one the predictions for students coming from the thirty schools with the lowest total means are too high. Since the number of students coming to the University of Wisconsin from these schools after excluding those from schools in which the variation from year to year is larger than would be expected from chance is so small this adjustment could not be made.

In conclusion the following statements can be made. In the sample of one hundred and eight Wisconsin high schools there is considerable variation in mean percentile score between high schools. The means for the senior classes over a five-year period vary between 33.1 and 68.2. The total means for the freshman classes in twenty Wisconsin colleges vary between 52.1 and 73.4. These variations are related to size of school as is indicated by a correlation of .44 between the total means and mean size of graduating class. In general the larger schools have the larger means. The need for an adjustment of ranks in graduating on the basis of these total means has been indicated. Four measures of variability within schools have been presented. These measures are: The correlation between means of successive classes, largest difference between means, average absolute deviations of the yearly means around the total means, and Fisher's z function. The correlation between

TABLE XII

ERROR IN PREDICTED GRADE POINT AVERAGES

	Number	Mean Percentile Score[2]	Mean High School Rank	Predicted Grade Point Average	Actual Grade Point Average	Error
High Group						
1929	354	59.37	60.15	1.152	1.184	-.032
1930	367	62.56	62.005	1.246	1.391	-.145
1931	283	64.04	62.52	1.284	1.347	-.063
1932	298	53.49	60.19	1.047	1.245	-.198
1933	270	54.74	64.18	1.147	1.217	-.070
Low Group						
1929	41	52.78	77.39	1.373	1.012	+.361
1930	52	52.41	74.22	1.304	1.142	+.162
1931	40	45.68	70.88	1.117	1.148	-.031
1932	51	53.91	76.76	1.381	1.094	+.287
1933	55	50.60	75.94	1.305	1.259	+.046

1. Wilhelm Reitz, "A Method of Adjusting School Marks and Class Ranks to Compensate for Variations in Standards of Marking", Bulletin of the American Association of Collegiate Registrars, IX (October, 1933), pp. 1-14.
2. These percentile ranks are those computed on the basis of college students rather than the state senior class population.

successive classes was found to range between .579 and .736. Corresponding figures for the college freshman classes are .84 and .78. The mean largest difference for the high schools is 12.21 and 9.72 for the colleges. There is a much wider range of largest differences for the high schools than for the colleges. The mean average absolute deviation for the high schools is 3.75 and 3.23 for the colleges. On the basis of these three measures college freshman classes vary less from year to year than do high school senior classes. Twenty-eight high schools and six colleges have variations between classes which would occur by chance in less than one out of one hundred cases.

FIVE YEARS OF EXPERIENCE WITH A REMEDIAL READING COURSE FOR COLLEGE STUDENTS[1]

by
J. R. Gerberich
University of Arkansas

The first re-training course in reading and methods of study at the University of Arkansas was introduced in 1928-29. During that year approximately forty probation students were given a course of this nature for eighteen weeks extending from the middle of the first semester to the middle of the second semester. Since that year approximately fifty freshmen whose reading abilities were found to be very deficient have been required to take remedial work during the first semester of each year.

Scores made by the freshmen on the reading and psychological examinations of orientation week were used as the major criteria in selection of students for the special course after 1929, with scores on the reading test as the more important of the two. Both the Study Performance Test, published by Ohio State University, and the American Council Psychological Examination were used with the entering classes of 1929 and 1930. The Ohio State University Psychological Examination has been used since 1930 to give a measure of general scholastic aptitude, while Part III of that examination, a reading test, has furnished a reading examination score. A few students who ranked exceptionally low on both the reading and psychological examinations were not included in the course because it was felt that their chances of success in college were almost hopeless. College entrants markedly over the average age for freshmen were excused and a few others were excluded for administrative reasons relating to their curricula, schedules, and the like.

From its inception, the special reading and methods of study course was given for two separate hours weekly without credit. The schedules of most of the students were reduced to thirteen or fourteen hours of credit work, although in a few cases the special course was superimposed on a normal schedule of fifteen or sixteen hours. The students were sectioned in two groups numbering approximately twenty-five each, thus permitting some attention to the problems of the individual during the class period. Conferences, analyses of results obtained from occasional tests, reports of scholastic standing, and other available types of information were utilized in an attempt to adapt the work to the needs of the individual.

Although the class procedure varied considerably, with new materials and methods introduced from year to year, the method of conducting the course remained relatively constant in the important elements. Varying types of reading exercises, accompanied by short check-tests, formed the major portion of classroom procedures. Discussion of the results of these exercises and comments by the instructor were interspersed with objective test exercises of standardized and unstandardized types. The units of the course dealt with such reading and study skills and abilities as word meaning, sentence meaning, paragraph comprehension, comprehension of longer selections, obtaining the main idea of selections, summarizing short thought units, outlining, taking reading and class notes, interpreting graphs, and increasing reading speed.

1.Dr. C. E. Prall, Dean of the College of Education of the University of Arkansas, has made many helpful suggestions and has contributed much of value in the preparation of this article.

By the use of a control group, it was possible to obtain objective evidence concerning the efficiency of the remedial reading course for probation students during the year 1928-29.[1] Probation students with scholastic and intelligence ratings comparable with those of students required to take the special reading work were selected for the control group. The two groups were compared on the following criteria: (1) final scores on an objective reading examination given at the close of the experiment, (2) scholastic standing at the close of the experiment and throughout the following school year, and (3) persistence in the University for more than a year following the close of the experiment.

Scores made by experimental and control students on the initial and final presentation of an objective reading test showed that the experimental students made twice as great a gain as the control students. Scores made by the two groups on an outlining and note-taking test, which made up a portion of the final measure, showed the experimental group to surpass the control group by a difference of 3.4 times the probable error.

The experimental students raised their scholastic standing 0.32 grade points[2] during the eighteen weeks of the experiment, while the control students lost 0.08 grade points during the same period. During the year following the experiment, students of the experimental group averaged 0.29 grade points higher than the control students, while carrying on average of 1.2 more hours of course work.

At the close of the experiment, 84 per cent of the experimental group and 65 per cent of the control group were still in school. Basing further persistence data on thirty-one experimental and thirty control students who completed the second semester of the year of the course, it was found that 42 per cent of the former and 30 per cent of the latter persisted long enough to enroll for the first semester of 1930-31.

In order to measure the efficiency of the remedial reading course given since 1929, it was necessary to establish some type of control group for comparison with the special reading group. The achievement measures used as indices of final attainment were general scholastic success for the first and subsequent semesters of attendance, percentages of students persisting varying numbers of semesters in attendance, percentages of students placed on probation because of poor scholarship, percentages of students dropped from their colleges of registration because of poor scholarship, and percentages of withdrawals during the progress of a semester. Table I presents evidence concerning the validity of scores on the reading and psychological examinations for use in establishing a control group to be compared with the experimental group.

TABLE I

COEFFICIENTS OF CORRELATION BETWEEN GRADE POINT AVERAGES FOR THE FIRST SEMESTER OF UNIVERSITY ATTENDANCE AND SCORES ON THE READING AND PSYCHOLOGICAL EXAMINATIONS

Year	G.P.A. vs. Reading Exam.*			G.P.A. vs. Psych. Exam.**		
	No. of Cases	r	P.E.	No. of Cases	r	P.E.
1929	293***	.47	.03	436	.46	.03
1930	254***	.62	.03	369	.58	.02
1931	392	.46	.03	392	.55	.02
1932	365	.56	.02	365	.62	.02

*The tests used were: 1929 and 1930 - the Study Performance Test, Form A, published by Ohio State University; 1931 and 1932 - Part III of the Ohio State University Psychological Examination.

**The tests used were: 1929 and 1930 - the American Council Psychological Examination; 1931 and 1932 - the Ohio State University Psychological Examination.

***As students entering the College of Engineering were not required to take the Study Performance Test, these correlations involve fewer cases than those based on the psychological examination.

The degrees of relationship indicated by the coefficients between the two initial measures and the most important of the final

1. The two sources of data which are briefly summarized here are as follows: C. E. Prall and G. N. Cade, "The Results of a Re-Training Experiment with Probation Students", Unpublished, 1929, 7 pages, and J. R. Gerberich, "Subsequent University Careers of Probation Students in the Re-Training Experiment of 1928-29", Unpublished, 1931, 5 pages.
2. A weighted grade point average based on the following system of weighting semester marks: A 6, B 4, C 2, D 0, E -1, and F -2.

criteria warranted the use of the reading and psychological test scores for the purpose indicated. The relationships between scholastic success and the other criteria of final achievement--persistence, probation status, dismissals for poor scholarship, and voluntary withdrawals--are all sufficiently close to warrant the use of the same equating factors for all final comparisons.

The students comprising the special reading class have been drawn from those ranking in the lowest twelve to fourteen per cent on the reading test and largely from the lowest fourth of the psychological test ranking. Since nearly all the students scoring in the lowest twelve to fourteen per cent on the reading test were required to take the remedial course, control students were not available for close equation with special reading students in the attempt to measure the efficiency of the remedial training. In lieu of exact controls, therefore, students were selected for the comparison whose scores on the reading test were just above the scores made by the students of the special course. Their rankings on the reading test were in the main from the twelfth to the twenty-third percentiles, although a few control students would originally have been included in the special reading group except for some difficulty in arranging the schedules or like matters. The ratio of men and women was kept constant for the two groups. Table II gives the status of the experimental and control groups on both the reading and psychological examinations.

The data of Table II show that the control group in all cases surpassed the experimental group in mean scores on both the reading and the psychological examinations. Furthermore, that the differences were statistically significant is shown by the fact that they were all more than four times their probable errors.

The usual procedure in a controlled group comparison is to look to differences in final achievement between two initially-equated groups for evidence concerning the relative efficiencies of the experimental and control techniques. As has been indicated above, a close matching of experimental and control students was not possible, so the two groups were established without any attempt to match individual students. Here a significant gap existed between the ability levels of the two groups. If, in the final measures, the disparity between the achievement of these groups is significantly less marked than the difference which existed between them in the beginning, it may well be assumed that the instructional efficiency of the experimental technique has been effectively demonstrated.

The two most significant measures of final status for this study are general scholastic success, as measured by grade point averages, and persistence in the University. The first of these is dealt with in Table III. Only those students who completed at least one semester of attendance or a sufficient portion of it to have final semester marks on record are included in the tabulations. The

TABLE II

INITIAL STATUS OF CONTROL AND EXPERIMENTAL GROUPS ON THE READING AND PSYCHOLOGICAL EXAMINATIONS

Examination	Year	Control		Experimental		Diff.	$P.E._{diff.}$	Diff. / $P.E._{diff.}$
		No.	Mean Score P.E.	No.	Mean Score P.E.			
Reading	1929	59	111.27 ± 1.46	48	86.15 ± 1.53	25.12	2.12	11.85
	1930	54	95.47 ± 1.73	45	71.34 ± 1.28	24.13	2.15	11.22
	1931	55	20.27 ± 0.29	48	15.83 ± 0.29	4.44	0.41	10.83
	1932	57	22.16 ± 0.30	47	15.75 ± 0.27	6.41	0.41	15.63
Psych.	1929	59	92.81 ± 2.44	48	76.38 ± 1.87	16.43	3.08	5.33
	1930	54	96.17 ± 3.68	45	65.50 ± 2.10	30.67	4.24	7.23
	1931	55	67.46 ± 1.59	48	52.19 ± 1.79	15.27	2.40	6.36
	1932	57	67.98 ± 1.52	47	47.24 ± 1.01	20.44	1.82	11.23

first semester data are based on the first semester of attendance for the 196 control and 165 experimental students who remained in the University at least that long; the second semester data are based on the 115 control and 92 experimental students who completed at least two semesters of attendance, etc. The decrease in number of students in attendance from the first to the seventh semesters is due both to student mortality and to the time limitation which shortened the attendance records for entrants of recent years. Thus the number of students as reported in the table should not be taken as an index of persistence.

Information concerning the persistence of experimental and control students by colleges of initial registration is given in Table IV. The students entering in each of the four years are again combined for the first, second, and subsequent semesters of attendance. The "Total" column in each section of the table shows the number of students who had sufficient time since entrance to put in the designated number of semesters in regular-session attendance. The final columns of each section give the percentages of students actually in attendance each number of semesters relative to the total number of those who could have attended that

TABLE III

SCHOLASTIC SUCCESS OF CONTROL AND EXPERIMENTAL STUDENTS IN RELATION TO PERSISTENCE IN THE UNIVERSITY

Semester of Attendance	Control		Experimental		Diff.	$P.E._{diff.}$	Diff. / $P.E._{diff.}$
	No.	Mean G.P.A. P.E.	No.	Mean G.P.A. P.E.			
First	196	1.09 ± 0.06	165	0.69 ± 0.07	0.40	0.09	4.44
Second	115	1.45 ± 0.08	92	1.16 ± 0.09	0.29	0.12	2.42
Third	80	1.50 ± 0.09	57	1.33 ± 0.12	0.17	0.15	1.13
Fourth	48	1.58 ± 0.11	30	1.20 ± 0.13	0.38	0.17	2.24
Fifth	33	1.53 ± 0.13	22	1.34 ± 0.15	0.19	0.20	0.95
Sixth	13	1.88 ± 0.17	14	1.75 ± 0.20	0.13	0.26	0.50
Seventh	9	2.33 ± 0.40	7	1.79 ± 0.26	0.54	0.48	1.13

An examination of the mean grade point averages of Table III shows that the control group ranks consistently higher than the experimental group. In only one instance--the first semester of attendance--is the difference a significant one, however. Therefore, it may be concluded that those control students who persisted more than one semester in attendance did not attain grade point averages differing from those of the comparable experimental students by a statistically significant amount. It is recognized, of course, that the two groups are not here paired exactly in terms of the data reported in Table II, for the withdrawals of some students had destroyed exact comparability. However, the same factors were in operation to affect withdrawal and persistence in the two groups, so that the conclusions drawn from the data of this table seem fair to each group.

period of time.

It is evident from the data of Table IV (see following page) that the experimental students remained to complete the first semester in a slightly greater percentage of cases than did the control students. Persistence is greater for the control group during the following four semesters. The maximum difference between percentages of the total remaining in attendance is the 10.2 per cent in favor of the control group for persistence of at least four semesters. For persistence beyond four semesters, the numbers of cases are hardly adequate as a basis for more than tentative conclusions, but it appears that what differences do exist are at least not unfavorable for the experimental group.

In Table V (see following page) are presented data concerning the last three criteria of final achievement mentioned above:

TABLE IV

PERSISTENCE IN UNIVERSITY ATTENDANCE FOR STUDENTS OF THE CONTROL AND EXPERIMENTAL GROUPS

Sem. of Attendance	Control			Experimental			Diff.*
	Total	In attendance	Per cent in attendance	Total	In attendance	Per cent in attendance	
First	225	196	87.1	188	165	87.8	-0.7
Second	168	115	68.5	141	92	65.2	3.3
Third	168	80	47.6	141	57	40.4	7.2
Fourth	113	48	42.5	93	30	32.3	10.2
Fifth	113	33	29.2	93	22	23.7	5.5
Sixth	59	13	22.0	48	14	29.2	-7.2
Seventh	59	9	15.3	48	7	14.6	0.7

*Positive differences show greater persistence for the control group and negative differences for the experimental group.

TABLE V

PROBATION AND WITHDRAWAL HISTORY FOR STUDENTS OF THE CONTROL AND EXPERIMENTAL GROUPS

Group	Control					Experimental				
	All Cases	First Sem.	Later	Total	Per Cent	All Cases	First Sem.	Later	Total	Per Cent
Placed on Probation	225	51	45	96	42.7	188	59	40	99	52.7
Dropped for Poor Scholarship	225	8	17	25	11.1	188	6	18	24	12.8
Withdrew Voluntarily	225	24	7	31	13.8	188	16	7	23	12.2

percentages of students placed on probation, percentages of students dropped for poor scholarship, and percentages of students withdrawing before the completion of a semester for which they had registered. The results are presented separately for the first semester of the freshman year and subsequent semesters of attendance, because scholastic difficulties and mortality are usually much more crucial and serious during the first semester than at any other time.

The probation data of Table V show that 43 per cent of the control group and 53 per cent of the experimental group were at one time or another placed on probation because of low scholarship. Many of the students so classified were subsequently removed from probation; others were dropped for poor scholarship, withdrew while on probation, or are still in attendance on probation. More than half of the probation students in each group were first placed on probation before the close of the first semester of attendance. Many of the remainder were placed in that status immediately after the close of the first semester.

Eleven per cent of the control group and 13 per cent of the experimental group were dropped from their college of initial registration because of poor scholarship. Some reentered later or gained admission to other colleges of the University. Less than half of this group were dropped before the close of the first semester, but many of the remainder were dropped so shortly after the close of the first semester that they were not allowed to register for the second semester.

Fourteen per cent of the control group and 12 per cent of the experimental group withdrew before the close of a semester for which they had registered and during which they had attended classes.

The differences shown by these data indicate that the control students encountered fewer scholastic difficulties than the

experimental students. However, in only one of the three comparisons--students placed on probation--is the difference of sufficient size to merit attention.

Summary

The purpose of this investigation was to obtain objective information concerning the efficiency of a remedial reading course required of approximately fifty University of Arkansas freshmen each year since 1929. The following factors were used as criteria of final achievement: (1) general scholastic success for the first and subsequent semesters of University attendance, (2) percentages of students persisting varying numbers of semesters in attendance, (3) percentages of students placed on probation because of poor scholarship, (4) percentages of students dropped from their colleges of registration because of poor scholarship, and (5) percentages of students withdrawing during the progress of a semester.

It was impossible to select a group of control students having abilities similar to the abilities of the experimental students on the two initial measures--scores on reading and psychological examinations. Therefore, control students were selected whose scores on these tests were in most cases just above the scores made by the experimental students. The differences in mean scores made by the two groups were found to be statistically significant and to show the control students to be distinctly superior on the two initial measures.

Differences in the standings of the experimental and control groups on the final criteria showed a tendency toward higher scholastic averages, greater persistence in the University, and few scholastic difficulties for the control students than for students in the experimental group. However, these differences were found to be relatively small and almost without exception lacking in statistical reliability. It becomes evident, then, that the control students, showing superiority of mean scores on two initial tests of determined validity, did not maintain their initial superiority on the criteria of final achievement. It may well be assumed, therefore, that the markedly less disparity existing between the final status of the two groups than that which existed on the initial factors is evidence of the instructional efficiency of the experimental technique.

INTERCORRELATIONS OF ARITHMETICAL ABILITIES

by
Leo J. Brueckner
University of Minnesota

In the Critique of the Twenty-ninth Yearbook of the National Society for the Study of Education it was proposed that four major functions of arithmetic instruction should be recognized, namely, the computational, the informational, the sociological, and the psychological functions.

If one accepts this view that arithmetic should be taught as a social study rather than as a tool subject, he is concerned with the devising of methods of evaluating the extent to which all of the desired objectives of the subject are being achieved.

In an attempt to make available a means of measuring the extent to which these functions are now being achieved, the writer devised a series of four scaled tests[1] dealing with several major aspects of these functions. One of the tests includes scaled exercises (examples in mixed processes) in computation; the second consists of scaled verbal problems of the conventional type; the third contains scaled exercises of the multiple choice type for measuring understanding of quantitative vocabulary, one aspect of the psychological function; the fourth test, called Understanding of Quantitative Relationships, includes scaled exercises of the multiple choice type for evaluating the ideas and the range of information pupils have concerning social institutions such as banking, money, taxation, business practices, applications of number, etc., that are emphasized by the informational and sociological functions of arithmetic. The last two tests did not contain any exercises involving computation, while all of the exercises in the first two tests did involve computation.

This paper presents a brief summary of the results secured when these tests were administered to a group of 453 pupils in grades 4A-5B in typical schools of St. Paul.

Table I shows the variability in arithmetic ability on each of the four tests for the whole group of children.

TABLE I

VARIABILITY IN ARITHMETIC ABILITY OF 453 PUPILS IN GRADES 4A-5B
(Expressed in C-Scores)

	Mean	Range	S.D.
Computation	63.9	37.5 to 91.5	8.6
Problem Solving	63.6	27 to 96	10.2
Vocabulary	64.5	24.5 to 103.5	10.8
Quantitative Relationships	65.6	23 to 92.5	9.4
P.E. of tests approximately	±2.0		

It will be observed that the mean scores on all four tests were practically the same. However, the range in scores on each of the tests was very large. Expressed in terms of grade differences the range between the highest and lowest scores is equivalent to more than five grades for each test.

The question naturally arises, what are the intercorrelations among the scores on the four tests? The answer to this question is given in Table II, on the following page.

The correlations between general intelligence and the various arithmetic test scores are all positive. They range from .350 between I.Q. and computation to .766 between I.Q. and vocabulary. The range of the skills measured by the test of computation is much narrower than the range of the learnings measured by the vocabulary test.

1. Analytical Scales of Attainment in Arithmetic. Educational Test Bureau, Inc., Minneapolis, Minn.

TABLE II[1]

INTERCORRELATIONS OF ARITHMETIC ABILITIES AND GENERAL INTELLIGENCE

	Computation	Problem Solving	Vocabulary	Quantitative Relationships
I.Q.[a]	.350	.506	.766	.602
Computation		.631	.361	.417
Problem Solving	.691		.522	.576
Vocabulary	.361	.522		.661
Quantitative Relationships	.417	.576	.661	

[a]Mental ability based on Von Wagenen Unit Scales of Aptitude; mean I.Q. of group .95.

The correlation between I.Q. and problem solving, a more complex ability than computation, is .506 and between I.Q. and understanding of quantitative relationships, a trait more complex than problem solving, .602. The correlations with general intelligence tend to increase with the increase in the complexity of the traits involved. The correlation between computational ability and vocabulary, .361, and between computational ability and understanding of quantitative relationships, .417, are both quite low. These correlations show that a pupil's rating on the computational test is a very unreliable index of his ability as measured by the vocabulary or quantitative relationships tests. Tests of each trait are essential for well rounded appraisal.

The general conclusion may be drawn that the more specific the traits involved, the lower are the correlations; the more complex the traits, the higher are the correlations. If tests were given to evaluate each of the specific outcomes that could be listed under an analysis of the four functions, it is probable that the correlations between many of the elements would be quite low, depending on the narrowness of the traits involved.

The results here reported show that for particular individuals there are marked differences in the levels to which the various traits have developed as a result of the learning experiences in which the pupils have participated. If the emphasis in instruction has been on the computational function we can be quite certain that the development of the traits included in the other functions will be uneven and unbalanced. To make certain that growth of all of the desired traits takes place, direct provision must be made for educative experiences in which they may function. Their development cannot be left to chance circumstances and cannot be regarded as a by-product of instruction which is at present largely limited to the development of skill in computation.

To get a picture of the variability of the abilities of bright and dull children as measured by these four tests, the results for two groups of children, one consisting of 117 children with I.Q.'s of 100 to 129, and the other consisting of 116 children with I.Q.'s of less than 85 were compared. Table III gives the means of the two groups on each of the four tests and the ranges in the scores.

TABLE III

VARIABILITY OF ARITHMETIC ABILITY OF BRIGHT AND DULL CHILDREN

	Means		Ranges		S.D.	
	Bright	Dull	Bright	Dull	Bright	Dull
Computation	68.7	58.4	55-86	37-76	6.61	5.79
Problem Solving	71.2	58.2	45-96	37-77	8.88	8.69
Vocabulary	74.4	55.4	515-98	24-71	8.35	9.36
Quantitative Relationships	73.4	58.5	49-90	23-79	7.06	10.36

The mean scores for the bright group in each case were larger than for the dull group. However, an analysis of the ranges for the two groups shows a wide overlapping of ability. In two tests, vocabulary and quantitative relationships, the dull group appears more variable than the bright, while in computation and problem solving the bright group appears more variable than the dull. The only significant difference in variability of the two groups is in the test of quantitative relationships where the dull group varies significantly more. The differences between the bright and dull groups in specific traits would undoubtedly

1. From an unpublished master's dissertation by Virgil Walker, "Interrelations of Arithmetic Abilities", on file at University of Minnesota Library.

be increased by intensive training; however, it seems quite clear that both groups would profit greatly from systematic instruction dealing with the different traits involved in the four major functions of arithmetic.

Judging from the overlapping of abilities of the two groups there is no reason to assume that the work in arithmetic for dull children need be limited to practice on computational processes as is now the common practice; the evidence is clear that they can profit greatly from participation in activities which deal with the social and cultural values of arithmetic.

The wide range in the abilities of these children makes it clear that instruction must be organized in such a way that the difficulty of the work may be adapted to the needs of the individuals. Even if pupils are grouped according to their mental ability the type of class organization for both superior and inferior groups must make full provision for individual differences in view of the wide range of the abilities and needs of the pupils in the two groups.

RESULTS OF USE OF MACHINES FOR TESTING AND FOR DRILL, UPON LEARNING IN EDUCATIONAL PSYCHOLOGY

by
James Kenneth Little
University of Wisconsin

The experimental study of problems of college teaching is a development quite largely of the last decade. However, these investigations have already yielded suggestions for the improvement of college instructional procedures which should, if effectively combined, markedly increase the efficiency of college work. A still newer development is the invention of various devices for automatically scoring tests and giving drill. Such devices should facilitate experimentation at the college level. Certain of them seem (as will be pointed out shortly) peculiarly suited for following out certain of the suggestions referred to above for the improvement of college work. The very practical problem of the investigation reported in this paper[1] was to see what might be accomplished, using certain of these testing and drill machines, to follow up some of these suggestions for the improvement of instructional efficiency.

Materials and Subjects

Previous investigations of college instructional problems had (the writer felt) emphasized the following relevant points: (1) the motivating effect upon the learner of knowledge of standing and progress; (2) the value, both for motivation and for guidance in learning, of informing students specifically and immediately of their errors and their successes in their work; (3) the value in the above (and other) connections of the frequent short test, as contrasted with less frequent longer tests or examinations; (4) the great importance of continuous adjustment to individual differences not only in capacity but also in error pattern and difficulty; (5) the value, in all these connections, of a consistent use of the make-up test.

Two pieces of apparatus available to the writer appeared especially useful in facilitating an instructional program developing the above points. The first was a device for the immediate automatic scoring of objective tests and tabulation of results by item.[2] This machine made it possible for an instructor to score a 30-question test sheet in less than a second. Immediately after he had finished a test a student could, therefore, know his score. And as soon as the last student's paper had been run through the machine the instructor could glance at the item-tabulation dials and at once discuss with the class those questions on which most mistakes had been made. The second or drill machine was a small device, one of which was given each student for use during the drill period. The student indicated his answer to each question by pressing one of five keys (one of two, for true-false questions). If the answer was correct the drum of the machine turned up the next question; if incorrect the same question remained before the student, and he was thus required to find the correct answer to each question before proceeding to the next. Meanwhile, a counter dial counted the total number of tries made in completing the test. The students were shown how to reset the machines, and were expected to go

1. The paper summarizes a thesis submitted in partial fulfillment of the requirements for the doctor's degree at the Ohio State University. The writer wishes to express his obligations to Dr. S. L. Pressey for his assistance in this investigation.
2. For a description see S. L. Pressey, "A Third and Fourth Contribution toward the Coming 'Industrial Revolution' in Education", School and Society, XXXVI (November, 1932). As used in this investigation the apparatus had been improved in certain respects, as in the use of hand punches for marking the answer slips, and in the addition of a mechanism which automatically printed each student's score on his paper.

through each test until they did so with no errors.[1]

The investigation was carried on during the winter and spring quarters in the 5 hour required quarter course in educational psychology, and dealt with the regular work of the course.[2] The course was taught in sections averaging 30 students, by five different instructors; in all, 14 sections took part in the experiment. The students were mostly freshmen and sophomores.

The course is divided into 14 teaching units; for each, two 30-item true-false tests were available. There were also used a 140-item pre-test (reliability .80) covering the entire course and given to all sections at the beginning of the quarter, a 100-item selective answer (5 choice) mid-term covering the first half of the course, a similar 100-item test covering the second half, and three broad essay type questions sampling from the entire course. In working up results the mid-term and final objective examinations were combined and considered one final test (reliability .92).

Procedure

A calendar was prepared at the beginning of each quarter. This required that the sections keep together, spending an average of about three days on each teaching unit, and that each test be given on the same day. Twelve such tests were scheduled, the last chapter of the first half of the course being included in the mid-term, and the last test of the second half in the objective final.

Each quarter, two sections of the course were run as "test machine groups", two as "drill machine groups", and the remaining three sections as control groups; there were thus in all four test sections, four drill sections, and six control sections. So far as possible, sections of each type were rotated among the instructors and times of day so that differences in instructor efficiency, time of day, etc., would be equalized; and careful examination of the data would indicate that such factors were largely negligible or thus equalized. The pre-test, and the University Intelligence Test (given to all entering students), made it possible to pair cases as to ability and preparation.

All sections took pre-test, mid-term, and finals. In the test machine sections each one of the twelve unit tests was scored by the scoring and tabulating machine; as each student finished punching his answer slip he brought it to the instructor, who at once scored it, checked wrong answers (very easily done when the slip is in the machine),[3] tabulated the score on a tabulation form already on the board, and returned the slip to the student. When all had finished, the test was discussed, the instructor directing the discussion toward the items which the item-tabulator revealed as most frequently missed. Each student thus at once knew his score. Since all the tests were 30 questions long he readily sensed how far he was from mastery of the material. From the distribution on the board he could see what his relative standing was; he could also see his letter-grade mark, since these were also indicated on the distribution. Further, all students who on the first unit test scored below B were required to take a make-up test (the duplicate form of the first test) on the following day; and this was scored and discussed in similar fashion. The final mark on the unit was then taken as the average of the first and make-up tests. The ways in which this procedure provided motivation by immediate knowledge of results and progress, informed of errors, and adjusted to individual difference, will be evident.

In the drill sections the same general procedure was used. But the tests were taken using the drill machines. The score the first time through was tabulated as

1. For a description of an early form of this machine see S. L. Pressey, "A Simple Apparatus which Gives Tests and Scores—and Teaches", School and Society, XXIII (March 20, 1926).
2. Various preliminary investigations were made, using nonsense syllables, German vocabulary, and unfamiliar Tartar history. They suggested progressively greater value of the procedures used according as the material was meaningful. Because of this, and because of the greater practical significance of experimenting with regular subject matter, the major effort of the experiment was therefore directed as above described.
3. A scheme is now being worked upon, for automatically marking rights, where a mistake has been made.

the student's score on that test; but the student then reset the machine and repeated the test until he could go through it without error. The drill sections thus had the advantages of the test section, plus review with automatically required corrections of error, and opportunity for immediate progress to mastery of that test.

The control sections took the chapter unit tests by marking on an answer slip which was graded by the instructor that evening and returned the next day. There were no make-up tests. The control sections thus did not have immediate knowledge of results (they did know results the next day, and the cards had errors checked so that they could locate mistakes). They did not have the stimulus of possible excuse from the make-up if they did well on the first test, the motivation involved in the possibility of improving their mark if they did poorly in the first test, or the help which taking the make-up and locating any errors might give them. Nor did they have the benefit of the diagnostic reviews with measure of progress provided by the drill machine. The control sections had this possible advantage, that the extra time taken by the testing and the make-ups, in the experimental sections, could in the control sections be devoted to teaching and class discussion.

Results

In handling results, each student in an experimental section was paired with a student in a control section on the double basis of score on the pre-test and on the intelligence test; pairing was fairly close (within five points) and in consequence the following data do not of course include all cases. The table on the following page shows the combined (winter and spring quarter) results for the "test machine" group and its control and the "drill machine " group and its control, on both the 200-item objective final tests and the 45-minute, 3 question essay final examination. Not only the medians, but also Q_1, Q_3, and the entire distributions are given since the differential effect of the experimental procedures upon the lower as compared with the upper part of the distributions is the most striking feature of the results. Thus it will be noticed that on the objective final test Q_1 of the drill group is 19 points above Q_1 of the control (7 points above the median of the control), whereas the experimental Q_3 is only 5 points above the corresponding control. On the objective test, the drill group shows greater gains than the test group. On the essay type final this difference also appears, but all differences are less. This is probably due in part to the unreliability of the grading of essay type questions. At the bottom of the table are shown the differences between experimental and control medians, the probable errors of these differences, and the differences divided by the probable errors. It should be again emphasized, however, that the most important differences are at Q_1 rather than at the medians, and that the results, and their reliabilities, are to be evaluated from an inspection of the entire distributions rather than from any one comparison.

Minor Exploratory Investigations

Certain special studies were made in an attempt to estimate roughly but more specifically the effect of some of the motivational and instructional techniques. During the spring quarter, one chapter test was repeated in all sections just before the make-up to see the effect of immediate discussion and drill in the experimental group on these same test items. The experimental groups gained one point only in the median but there were considerably fewer very low scores.

A test compounded from the most difficult items of tests occurring early in the course, and another similarly compounded from the later tests, were given to see whether the experimental groups had especially benefited, as indicated by improvement on these questions. Results were not entirely consistent, but gains were greatest in the experimental groups.

Study was made of the effect of the

TABLE I

PAIRED CASE RESULTS, SPRING AND WINTER QUARTERS COMBINED

End of quarter differences between (1) test machine group and its control, and (2) drill machine group and its control, cases being paired on the basis of (a) intelligence and (b) a pre-test on the subject matter of the course.

200-item multiple choice final test							A 45-minute, 3 question essay type final examination						
	TEST			DRILL				TEST			DRILL		
	Exp.	Con.		Exp.	Con.			Exp.	Con.		Exp.	Con.	
185-9	1			1			60-2						
180-4	1	1		3	1		57-9				1		
175-9	1	2		2	1		54-6	4	1		1		
170-4	8	7		4	3		51-3	1	2		2	1	
165-9	8	6		6	4		48-0	5	7		4		
160-4	11	3		7	5		45-7	5	4		5	4	
155-9	9	6	Q_3	9	8	Q_3	42-4	13	3		11	3	
150-4	10	5		8	5		39-1	7	11	Q_3	7	6	
145-9	10	13		10	1		36-8	10	11		7	8	Q_3
140-4	12	8	M	6	4		33-5	10	9		2	12	
135-9	6	7		3	4	M	30-2	11	9	M	14	9	M
130-4	6	11		3	8		27-9	14	10		4	5	
125-9	9	9	Q_1	2	8		24-6	5	11	Q_1	4	5	Q_1
120-4	4	5		3	6	Q_1	21-3	4	7		3	5	
115-9	1	4		3	3		18-0	6	5		3	3	
110-4	1	5		1	3		15-7		7		1	7	
105-9		4			4		12-4		1		2	3	
100-4		3			2		9-11	1	1				
95-9					1		6-8	3					
90-4							3-5						
85-9	1						0-2						
N =	99	99	diff	71	71	diff	N =	99	99	diff	71	71	diff
Q_3=	163	156	7	163	158	5	Q_3=	43	40	3	44	38	6
M =	150	141	9	153	136	17	M =	35	32	3	37	32	5
Q_1=	138	127	11	143	124	19	Q_1=	28	25	3	30	24	6
ME - MC	9			17				3			5		
PE of Diff	1.92			2.34				1.06			1.17		
Diff/PE	4.69			7.26				2.83			4.28		

(ME = median of experimental group; MC = median of control group.)

three trials by drill machine; it might be supposed that students having once been apprised of their specific errors, and having access to the textbook for reference, would on a second or third trial be able to answer a test with a perfect score. However, median scores of a class on three trials on a representative chapter test were 23, 27, and 29. The range of scores progressively narrowed and the scores were bunched at the top on the third trial; but the slowness with which this improvement took place, in view of the opportunities at hand, was rather striking. Students missed items on the second and third trials which they answered correctly on the first. The results are an interesting suggestion as to the probable slow and "developmental" character of learning.

An experimental section and a control section, both taught by the same instructor and roughly equal in intelligence and previous knowledge of the course, were compared in terms of the grades received on the first test given on each chapter. Presumably, neither of the experimental groups would have the advantage of the special experimental procedure of reviews and diagnosis in such a comparison, but more general motivating effects of the total procedure (plus cumulative gains in preparation for each succeeding unit of work)

might appear. It was found that approximately twice as great a percentage of the students in the control group received D's and E's as in the experimental group, and fewer made A's and B's. The same general results obtained in comparisons of other such classes.

Questionnaires designed to secure student criticism and comment, upon the procedures used, were given in all sections. In general, the students reported that they preferred to take tests by the methods used rather than by the usual methods; they considered make-up tests desirable and indicated that as a general practice they studied chapters further in preparation for them. They voted more than 4 to 1 that there was not too much testing, and mentioned frequent tests as one of the most desirable features of the procedure. Only one student reported that he had ever taken a course before in which his test results were regularly made known to him and discussed on the same day as the tests were given. They reported their marks to average about the same, if not a little poorer, in the course than in other five hour courses, although they thought the course was somewhat easier. Cheating was reported, but few reported that it was more than is usual in college classes. They voted 9 to 1 that their total reaction to the procedure was quite favorable. The most undesirable features of the procedure, in the student's estimation, were tricky, ambiguous test questions, and too high standards for grades, particularly for A's and B's. Control sections complained that they should have had make-up tests too. A few in the drill machine group were disconcerted by machines which operated faultily, and a very few were annoyed by the noise of the machines. Some thought that there should have been more essay tests, and some asked for more time for class discussion.

The Larger Significance of the Study

The investigation here reported was planned as a very practical class experiment. A variety of factors which might presumably contribute to increased efficiency were therefore included; and analysis of the contribution of each is hardly possible. However, two general conclusions of very considerable importance may, the writer feels, be drawn. In the first place, the experiment is (so far as the writer knows) the first instance of the systematic use of machines as part of the regular routine in regular university classes. It demonstrates (he feels) the possibility of mechanical aids in education.

In the second place, the experiment is in interesting contrast to present trends in efforts to improve college instruction. The present trend seems to be toward the expensive tutorial system, toward still more expensive housing plans, and toward somewhat transcendent or sentimental emphasis upon the importance of "personality" in teachers. In contrast, one purpose of this experiment has been to show that an impersonal attempt to organize procedures in college instruction for greater efficiency may produce demonstrable results and at the same time save labor. In this time of depression such a saving may be of some interest.

Conclusions

(1) Students immediately apprised of their test results, and given opportunity to correct deficiencies by make-up tests, profit markedly in terms of final examination results over students who do not have such advantage.

(2) Students immediately apprised of the correctness or incorrectness of their responses to each item of a test, and given opportunity to correct deficiencies by drill and by make-up tests, likewise so profit.

(3) The greatest benefit accrues to students who usually score in the lower half of the distribution, although the entire groups move upward.

(4) Mechanical test-scoring and drill devices have a practical use in the classroom. They are convenient for students, time- and labor-saving for teachers, and make possible instructional techniques not otherwise practicable.

ELEMENTS OF CHORD-READING IN MUSIC NOTATION

by
Otto Ortmann
Department of Research
Peabody Conservatory of Music

The following test was undertaken as an initial step in a comprehensive study of the problems of music reading. In no way is it intended as a measure of music reading ability. It merely isolates one phase of such reading and attempts a diagnosis of the principles underlying this phase.
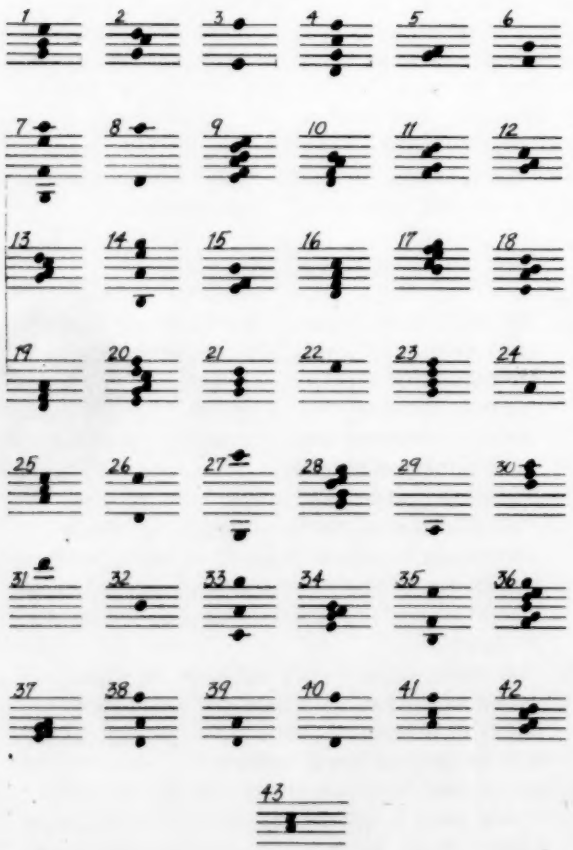
Figure 1

The element selected was the vertical grouping of notes as this is found in various chord-forms. A series of "flash-cards" (Fig. 1) was used in classroom presentation. The size of these cards was determined by approximating the ratio between note-size and eye-distance found under normal music-reading conditions. (The normal distance of the eye from the note-rack varies between 20 and 30 inches.) The staff lines used were 7/8" apart, the long diameter of the elliptical note-heads was 1". Preliminary testing indicated that for normal vision this size was sufficiently large to present a clear visual stimulus for classroom use. In order to eliminate all factors other than the spatial distribution of notes themselves, clefs and stems were omitted and black note-heads used. At a "ready" signal each card was exposed for 2/5 seconds under adequate light conditions. The pupils, immediately after each presentation, wrote on staff paper "what they had seen." After completion of the test, they were asked to comment, so far as possible, on the methods used and on any particular points of ease or difficulty. A tabulation of the replies revealed the presence of definite measures of approach and also of particular visual difficulties.

Number of Notes

The functioning of visual-span in reading, or, for that matter, in any visual perception, is well known. Consequently, its operation in the reading of music symbols may be expected. The average number of errors in 27 papers for one-note examples was 0; for two-note examples, 2; for three-note examples, 3.4; for four-note examples, 13.7; for five-note examples, 27.5; for six-note examples, 42.3. For the pupils tested, all

advanced students in music, reading difficulty, other things equal, begins with the four-note examples. The steady increase of difficulty with an increase in number of notes is shown in Fig. 2.

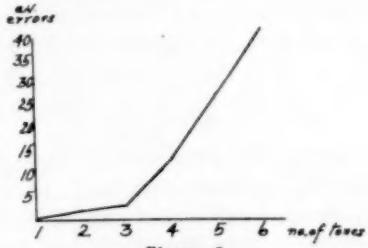

Figure 2

The actual number of notes is seldom missed. Many of the omissions found for certain examples are the result of failure to write a note when in doubt as to its exact staff-location and not from failure to read correctly the total number of notes present in the stimulus. "Sometimes I didn't put down any note because I wasn't sure just what note it was", indicates this reaction. This was noted by several students.

Six-note groups were found difficult (with the given exposure time) for even experienced sight readers of decidedly superior ability. Since, in a moderate Allegro tempo, more than six notes are very often correctly read in 2/5 seconds, it follows that some other form of grouping than the note-reading must function. Indications of this are given in later paragraphs.

Vertical Distance

A second determinant of eye-span is the spatial area. Since, in the present test, only the vertical dimension was used, this area determinant can be measured by vertical distance. If we array the test data on this basis, in a gross way, we get the following distribution for average errors:

	2-note	3-note	4-note	5-note	6-note
Vertical distance of a fifth or less	0	1.9	4.5	--	--
" " between 5th and octave .	0	1.0	10.5	27.5	30.
" " more than octave	3.6	9.7	30.7	--	38.5

The presence of other factors, yet to be considered, obscures somewhat the correlation between vertical distance and difficulty of reading, yet the given table indicates a clearly marked general increase. If we select examples, such as Models 10, 13, 18 and similar discords, the correlation between vertical distance and number of reading errors is marked. Using 4-note examples:

Vertical distance ..	4th	5th	6th	7th
Average errors	0	8	9	13

Vertical distance ..	10th	13th	16th
Average errors	22	28	42

Again, if we take similar chord-forms used in both close-position (small vertical distance) and open-position (wide vertical distance) the greater number of errors is always found for the greater distance. Thus errors for M. 38 were three times as frequent as those for M. 19.

By restricting the count to specific intervals of various models having similar outlines, such as: Models 37, 42, 11, the errors are respectively: 0, 8, 14; Ms. 34, 10, 18 with errors 5, 12, 19. Both series indicate an increase of error with an increase of vertical distance, although the factor of chord-unit, discussed later, also affects this distribution.

Additional proof of the distance difficulty is given by the introspections of the pupils, fifteen percent of whom wrote: "wide intervals are harder than close ones", volunteering the information without any leading question from the examiner.

Expansion and contraction. The vertical distance of the visual pattern also affects the frequency with which the distance is overestimated (expansion) and underestimated (contraction). Measured by the extreme notes (total vertical distance) this was overestimated in 30% and underestimated in 70% of the total number of

examples. In the 4-note examples embracing more than an octave the errors were: exp. 27%, contr. 73%; for those less than an octave 44% and 56%. This indicates a general tendency to underestimate the vertical distance, the tendency increasing with an increase in vertical distance. In examples such as Ms. 7 and 27 it is very pronounced, the former was contracted in 89% of cases, the latter in 94%.

The effect of vertical distance on the overestimating or underestimating of patterns, however, is also somewhat modified by the particular note distribution. Intervals, such as sevenths may be written as octaves because of the familiarity with the harmonic significance of the latter. An open position, such as M. 38 may be recognized as a triad and thus written correctly in spite of the wide vertical distance.

Form of Visual Pattern

Notes written in standard staff notation necessarily take on various patterns as a result of the spatial distribution of the notes. Such patterns are illustrated by the models given in Fig. 1. Examples like M. 2, 5, 9, 10, etc., not only have a vertical dimension but a horizontal dimension and, consequently, they best serve to illustrate the extent to which pattern-outline is observed in reading notes. The instructions of the test did not mention pattern, the pupils were free to write the notes in any spatial relationship desired. A number of combinations for such examples as 9 or 17 are possible and all would be correct from a notation standpoint.

For the single vertical lines (such as M. 1, 3, 7, etc.) the form was retained in 100% of replies; for the mixed groups (oblique relationship or two verticals) the correct form was retained in 90% of replies.

More significant than this, as indicating the importance of visual pattern in note-reading, is the fact that the geometric outline was retained even when reading errors occurred. The following illustrations are typical; the large notes are the models, the small notes are the written versions:
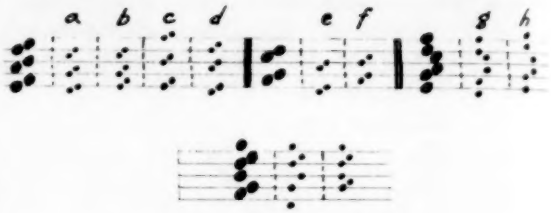
Figure 5

They show, in some cases, complete retention of the pattern outline; in others, retention of the vertical and oblique aspects with errors in the details. The visual pattern is correctly reproduced but the relationship between individual notes and the staff is incorrect.

If this be true, then the simpler patterns should be easier to read. Simplicity or complexity is determined by the linear relationships. When the notes of an example appear to form parallel lines in both dimensions as in Models 11 and 42, the group is more readily perceived than when they do not form parallel lines in both dimensions, as in Ms. 10 and 13. Several examples permit a comparison on this basis. For the 4-note examples:

Parallel structure Ms. 11, 37, 42;
average error 7.3
Mixed structure Ms. 10, 18, 34;
average error 10.3

M.9, on account of the number of tones present, was a difficult example with many errors, yet the two verticals and three ascending obliques (left to right) were changed in only 3.7% of cases.

Further evidence of the significance of the visual pattern is found in the introspections of the pupils: "I saw the notes as a design or pattern." "Names of notes not thought of, worked on pictures of the groups presented." "Shape of chord-forms was helpful." "Some formed single vertical line, others two diagonals, still others had one note on a second vertical line."

Since all visual stimuli are grouped on the basis of proximity in space, the patterns used will necessarily show

subdivision into smaller groups according to the manner in which the notes are distributed. Examples, such as Fig. 4a, are not subdivided on the basis of spatial relationship; the remaining examples of Fig. 4 are so subdivided, the grouping into two units increasing from b to f. This

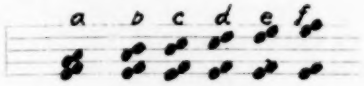

Figure 4

spatial grouping, which is a basic visual reaction, is interfered with by the vertical reading (two fourths or two fifths, etc.) resulting from the purely vertical construction of all chord-forms in music. Both types of reading were used by the pupils. Vertical type: "Noted whether intervals were seconds, thirds, etc. Also watched for octaves." "Even in No. 11, I read two perfect fifths." Spatial type: "Some chords, such as No. 11, divided themselves into two close groups, seconds." "In No. 11,I divided it C going to D, and F to G."

In examples of only one vertical line, notes will also be grouped through spatial proximity. Fig. 5 shows the distribution of errors for the given examples: large, black notes are the stimuli, small, black notes the incorrectly written notes; small, white notes are the omissions.
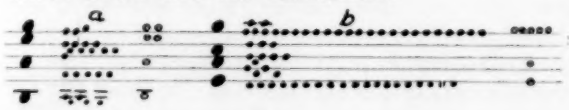
Figure 5

The eye is at first attracted to the portion of the pattern with closest notes; at a this is the upper third, at b it is the third in the middle. As a result, there are fewer errors for these parts than for the remaining parts. This is especially noticeable in b. In M. 14 two pupils wrote: "The top two notes grouped themselves together." Another general introspection was: "First saw centre of group where most notes were together." (Example 14 is an instance in which the ascending reading is interfered with by the high location of the visual centre of gravity.) For the same reason, we find, as typical variants of the following example, such distributions as in Fig. 6.

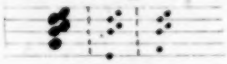

Figure 6

The "togetherness" of notes applies chiefly to seconds and thirds, since, in both cases, the notes are adjacent, hence without open space between. With the interval of a fourth, the two notes comprising the interval are separated by a white space, hence the grouping is less strong. The degree to which it then functions is a relative matter, depending upon the distribution of the other notes of the chord. In M. 7 the upper two notes are grouped because the interval of a fourth is the smallest interval in that example. In M. 18 the two lower notes (also a fourth apart) are not grouped because the remaining intervals are smaller. Grouping, therefore, varies with each chord.

This variation does not hold for the more fundamental grouping of seconds and thirds into "close" and all other intervals into "open", which remains the basic division for notation.

However, if an open fifth is "filled in" with an additional third, the three notes become a unit. Compare, in this respect, M. 39 with M. 19. For a similar reason, the interval of a seventh, open in M. 14, may become "close" by filling in two thirds, as in M. 16. With such additions a visual centre appears, whereas in the two-note open intervals there is no visual centre; they remain visually ungrouped. When the distance is "filled in" the visual centre is along the middle of the chord.

The ease, therefore, with which vertical notes are grouped, depends upon the spatial distribution.

Number of Verticals. In their remarks several pupils noted that: "Examples with more than one vertical line are difficult." This difficulty occurs when the verticals do not agree with familiar chord-forms. M. 9, read vertically, is difficult on this account; read obliquely, as three seconds, it is considerably easier. In such cases the model is frequently changed into familiar chord-units as in Fig. 7 written for
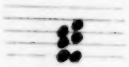
Figure 7

M. 9. And, conversely, when the verticals are familiar chord-forms, the vertical reading is advantageous.

In other examples the outline suggests definite grouping. In M. 20 the grouping is three thirds. Yet if this example be written as in Fig. 8, any version of which is correct notation, the division into three equal groups is entirely lost, and a

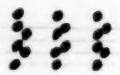

Figure 8

complex grouping results. Thus the difficulty of the example is increased.

It follows that when chord-groups are printed they should contain as many repeated elements as possible (Fig. 12a on page 56) which are somewhat easier than the forms shown in Fig. 12b on page 56. Sometimes, an even arrangement is impracticable, but within certain limits, it can be done.

The test results show a general tendency to read from the bottom note up. In Fig. 9 the error and omission distributions are given for various examples, including a few widely dispersed examples. These are selected because the wide vertical distances will reveal more clearly the direction of reading. If this be from the bottom note up, the greatest number of errors, including notes and especially omitted notes, will be found among the upper notes. (See Mo. 7, 20 and 36.)
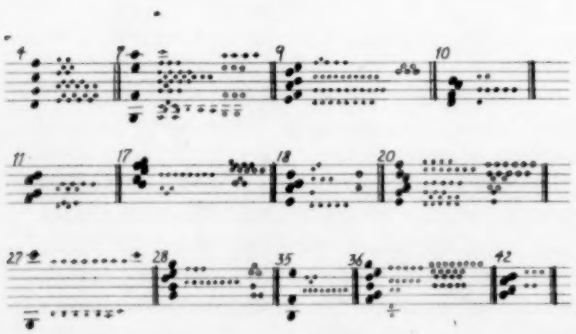
Figure 9

To the question "Did you read all the notes at once, or did you read from the top down or from the bottom up", 96% replied, unequivocally "from the bottom up", modifying it with the remark that for close groups, such as Nos. 10, 42, etc., the direction of reading was not noticeable.

This preference for ascending reading may readily result from associations with the construction of all chord-forms in music, which is invariably in an ascending direction. Since all the pupils taking the test were thoroughly experienced in the writing of chord-forms, the association, naturally, was strongly developed. On the other hand, it does not follow that the ascending reading, per se, has any advantage over a descending reading. The obvious advantage accrues from the readiness with which it fits into other notation experiences. In the reading of Ex. 9, for instance, the reaction "two ascending adjacent pairs of fourths on E - F", would involve concepts very familiar to the student and would include all the relationships among the notes of this example. Whereas, the reaction: "Six-notes, the bottom one on bottom line, next one on bottom space, next one on second space, next one on third line, etc.", obviously is so complex as to preclude its functioning effectively in the short exposure time. It is, therefore, natural to employ the interval and notation concepts already established, and since these are all ascending in nature, they explain the ascending reading. In view of the number and strength of these associations, it is probably wise to encourage ascending reading independent of the particular clef used.

However, as mentioned under Visual Centre, a close grouping of notes at any level, fixes the eye at this level. In Ex. 18 the note most often incorrectly read is the lowest. The omissions still point to an ascending reading. The direction, therefore, seems to remain, but the visual centre focuses the eye and draws it away

from the lowest note before this has been adequately read. When it is then written, guess substitutes the octave of the highest note, a combination found more frequently than the given example. This is borne out by the introspections: "watched for octaves", "it hurt to write intervals such as ninths, and I always wondered whether it was an octave instead."

Chord Units

In some ways the most important element in chord reading is the higher-unit grouping on the basis of chord-construction. That is to say, the recognition of a note-group as a single chord-unit, regardless of the actual number of notes involved. Chords are constructed in thirds. The smallest form is the triad, consisting of two thirds, one immediately above the other. Since thirds are "close" intervals in notation (notes touching each other) patterns such as Models 19, 21, 25, 30 are units both from harmonic construction and from visual grouping. In such cases, the theoretical and the visual reactions coincide, with the result that such examples are the easiest to read. Not a single error in reading "triads" occurred in the test. The ease of reading is independent of the exact position of the group on the staff.

The next larger chord-unit is the seventh-chord, which consists of three ascending thirds (Ms. 16, 23). Here again the theoretical and visual reactions coincide, and the frequency of reading error is again low: 3%.

A slightly less familiar chord-unit (first inversion of a triad) is given in Ms. 1, 41. When these examples are compared with other 3-note examples which are not chord-forms, such as Ms. 2, 12, 15, the ratio of reading errors was found to be as 3% to 15%. When the 4-note chord-forms with their 3% reading error are compared with M. 18, which has the same vertical distance and the same number of tones,(but is not a chord-unit) with its 60% error, the effect of chord-unit upon ease of note-reading is clearly revealed. It is the musical parallel of word reading as against syllable or letter reading.

The use of this higher-unit device is reflected in the introspections: "Read the notes as chords; hard when this was not possible." "Saw those with regular forms as a whole: V7, II7, etc." "Found triads and their inversions and seventh chords easier than non-chord forms." "Read as chords whenever possible." "In No. 14 saw upper three notes as incomplete 7th chord." "G-B-D was a triad."

When the chord-unit and the visual unit do not coincide, as in Ms. 13, 34, the reading is of moderate difficulty (38% error). In these examples the visual unit is clearly marked, but it does not suggest a chord-unit.

When neither visual-unit nor chord-unit is present, as in Ms. 4, 7, 14, the difficulty in reading increases further (74% error). Examples with wide vertical dispersion and non-chord units, form the most difficult examples. The difficulty begins to appear even with 3-note examples.

Attempts to substitute chord-units for a mixed visual pattern are indicated in a few reactions to M.36 which was grouped F-B-D and G-E-G in spite of the 4-2 vertical grouping of the pattern. The tendency to think in familiar chord-units is reflected also in the type of errors made. Thus, Ms. 9 and 20 were typically written as at a, d, g, h Fig. 3, showing the familiar octave substitution for the sevenths.

A typical reaction to Model 10 was to write the upper two notes a step lower. This may result from the general tendency to "contract" wide intervals, but may also result from the sbustitution of the familiar triad reaction, which then forces the note on the right into a wrong position in order to retain the pattern outline.

When, in a preliminary test, the notes C-F-B were given, the following error distribution occurred, showing clearly the substitution of a familiar note (a) for the less familiar (b) (Fig. 10).

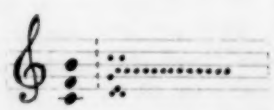

Figure 10

Space and Line Relationship

The fundamental division of notes on the staff is into notes on lines and notes in spaces. When the errors are tabulated on this basis we get the following ratio: lines written as spaces 81 times; spaces written as lines 106. For the same examples, the frequency of lines to spaces is as 28 to 36. Assuming that the number of errors is proportional to the frequency of lines and spaces we have the following proportion 81 : 106 :: 28 : 36 or 2916 = 2948 which is close enough to indicate an equal distribution of the error. In other words, lines are read as spaces about as often as spaces are read as lines.

There are two exceptions to this rule. First, when the example contains only spaces, the error of reading a space for the adjacent space exceeds the error of reading the space for the adjacent lines. In other words, the entire pattern is moved up or down to the next space. Second, when only leger lines are used (as in M. 27, the errors of reading line for adjacent line and space for adjacent space exceed the errors line-space or space-line. (See Fig. 9, M. 27.)

Taken collectively, however, the error of one staff degree (line-space, space-line) exceeds the error of two staff degrees (line-line or space-space). The most frequent error made is the misreading of a note by one step.

When symmetrical points of a pattern have the same line and space relationships the example is easier than when these relationships are absent. Thus Fig. 11a is easier than Fig. 11b.

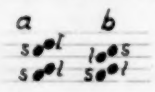

Figure 11

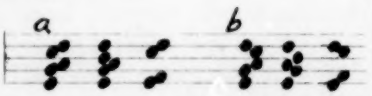

Figure 12

This, in turn, makes thirds, fifths and sevenths somewhat easier than fourths and sixths, when allowance is made for differences in vertical distance. And, conversely, when lines and spaces are present in mixed order, as in M. 9, the mixplacement of lines and spaces becomes a typical error. In one instance the entire relationship was reversed (see Fig. 3a).

In this example, as well as in Ex. 17, many partial reversals of lines and spaces occurred, since the stimuli themselves present a mixed outline.

A slight tendency to substitute one space for another (instead of an adjacent line) is shown in Fig. 5a and Fig. 9, M. 35. In both examples a single note is rather widely separated from its neighbors. Consequently, its space-line position is more readily seen, but the exact position of the space or line is no clearer than when the notes are grouped. Hence, the error is one of number of space or line rather than one of space-line or line-space. The converse of this is given by the error distribution for M. 42. Here, where the notes are closely grouped the space-space or line-line error is entirely absent. Similar instances are Ms. 9, 17, 18, 36. The fact that the most frequent error is that of one staff degree indicates that the spatial dimensions (vertical distance) is noted more readily than the line-space position of a single note. This is the typically increment error common to all sensory discriminations: small increments are more readily misread than larger increments.

Mental Set

If a chord-unit be anticipated and the stimulus by chance coincides with the anticipated unit, the example becomes easy. Thus M. 20 can be divided into a 4/3 chord on the lowest note, plus the top third, or as a 4/3 on (if treble clef, a) with the lower third. The errors for the two types of reading will differ since the visual focus will differ. It is a phenomenon similar to the staircase illusion and others, in which the mental set varies the grouping of the visual stimuli. In actual music reading this mental setting plays an exceedingly important role. The data secured in this test merely indicate its presence. One pupil wrote: "I find myself expecting certain chords; when I'm right the reading of even complex patterns becomes easy."

The mental set need not be complete before the stimulus is presented. If, in ascending reading, the two lower notes suggest a familiar chord-unit, the mental set is established and the remaining notes are

typically forced into the expected chord-form. This has been experimentally shown in some of the error distributions of Fig.3.

Note-Names

Although, in the test as given, no clef was used, hence the notes could have no names but were merely positions on the staff, yet in practically all cases names were assigned. In 70% of the papers the treble clef was definitely mentioned; an additional 15% indicated by chord-forms (V7, II7, etc.) that they were reading in the treble clef. The remaining 15% made no mention of what, if any, clef was being used. One paper read: "Names of notes seldom thought of, rather worked on pictures of the groups presented." Another read: "In all simpler forms I heard the sound of the notes"; this means that although note names may not have been present to consciousness, the pupil had at least assigned relatively definite pitches.

It does not follow, however, in reading complex notation stimuli and immediately reproducing them on a keyboard, that individual notes are read. In fact, the tendency to read in higher-units, described under Chord-Units, shows clearly a reduction of the separate notes into a note-whole. But the data show that note-names are more concrete than memory in terms merely of line and space relationships. This condition pertains, of course, to any group of music students; non-music students would necessarily have to rely upon the geometric pattern, the line-space distribution.

Sight Reading

The data of this test cannot be used as indices for what is generally called sight-reading ability. This includes reproduction, either vocally, or on some other instrument, of the visual stimulus. Since the test did not involve such reproduction, it cannot give a complete sight-reading index. The ability quickly and correctly to see the visual stimulus does not include the abilities to reproduce it. (However, the ability to reproduce it does include the ability to read it quickly and correctly.) The pupil making the second highest score in the notation test (95%) is conspicuously defective in reproducing visual stimuli on the piano, in spite of an adequate technical (kinesthetic) equipment. The pupil rating third highest (94%) is below average in sight reading. The pupils rating first, fourth and fifth are excellent sight readers. Other pupils making fairly high scores (between 80 and 90) vary in sight-reading ability from good to very poor.

Such a distribution shows that when visual stimuli are presented in their vertical aspect alone they are sufficiently dissociated from the visual stimuli as used in actual musical compositions to include only one element of the sight-reading process. And it shows, conversely, that sight reading is a complex process which does not depend upon the perception of single note-elements.

On the other hand, the principles underlying the reading of the stimuli as used in this test probably function in the reading of complex stimuli as well.